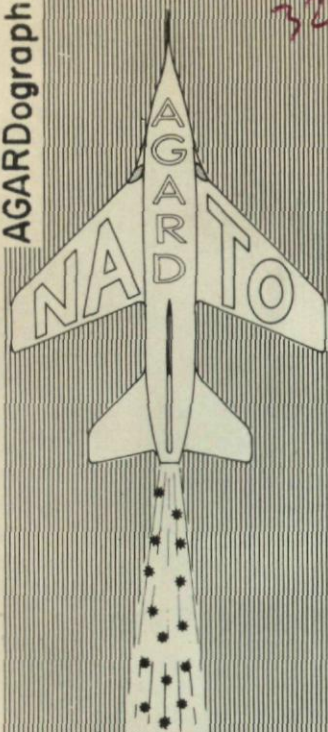


32 L63588 (B).

cm

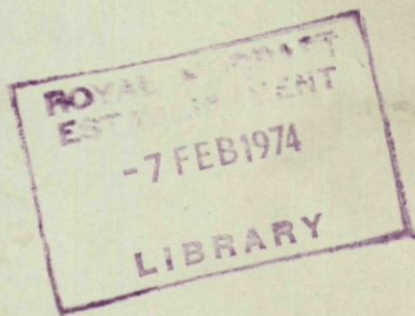


AGARD

SECOND GUIDED MISSILES SEMINAR GUIDANCE AND CONTROL

BELGIQUE
*
CANADA
*
DANMARK
*
DEUTSCHLAND
*
ELLÁS
*
FRANCE
*
ISLAND
*
ITALIA
*
LUXEMBOURG
*
NEDERLAND
*
NORGE
*
PORTUGAL
*
TURKIYE
*
UNITED KINGDOM
*
UNITED STATES
*

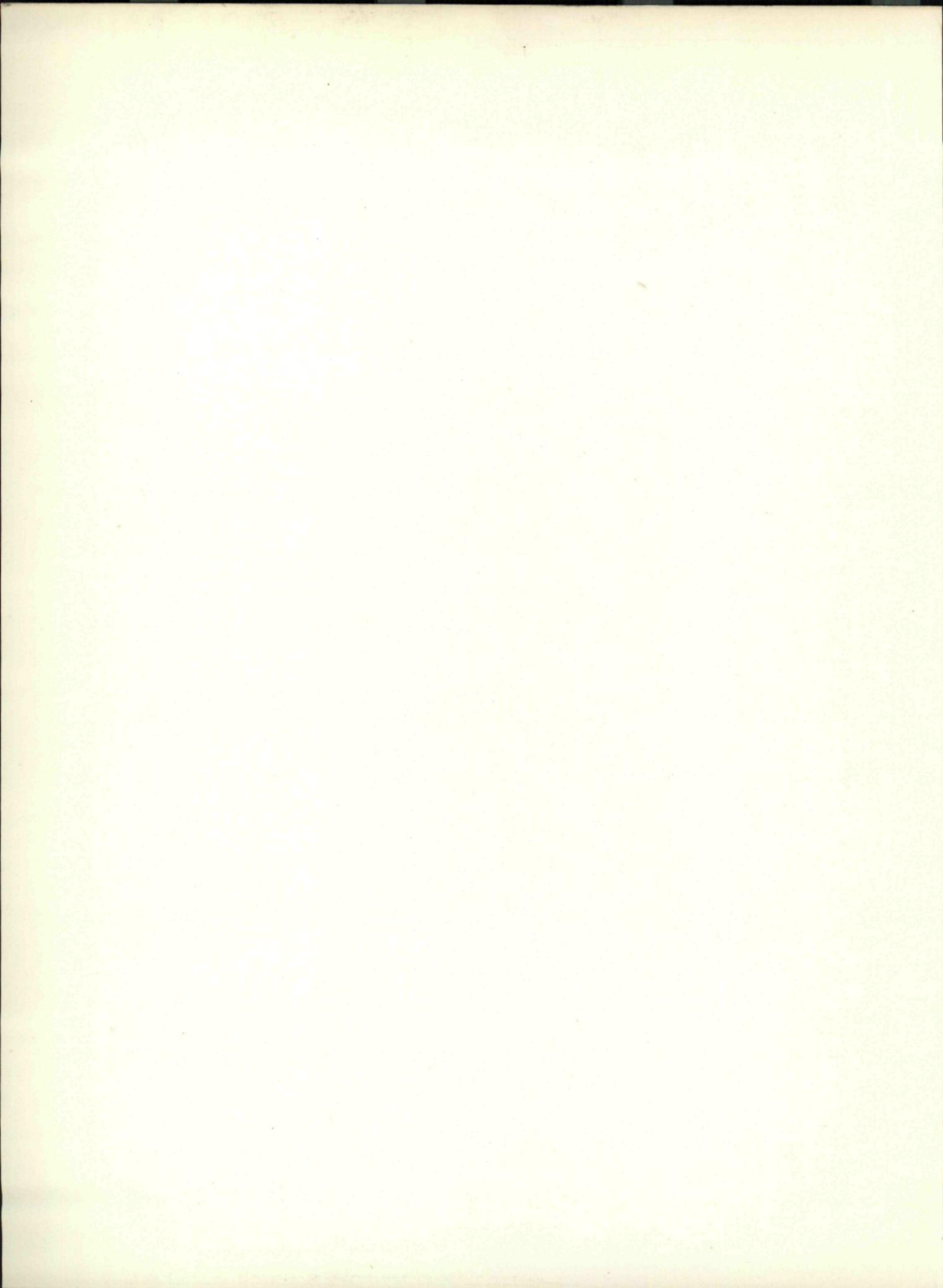
*space
G.W. Hilborn*



VENICE, ITALY
SEPTEMBER 1956

7705 B. (Book two)

<p>North Atlantic Treaty Organization Advisory Group for Aeronautical Research and Development</p> <p>SECOND GUIDED MISSILES SEMINAR GUIDANCE AND CONTROL 24-28 September 1956, Venice, Italy</p> <p>Twenty-four papers by different authors are included. The main topic dealt with is the guidance and control of missiles. Specific papers consider the range of problems from the selection of a missile guidance and control system to its final testing in actual flight.</p>		<p>North Atlantic Treaty Organization Advisory Group for Aeronautical Research and Development</p> <p>SECOND GUIDED MISSILES SEMINAR GUIDANCE AND CONTROL 24-28 September 1956, Venice, Italy</p> <p>Twenty-four papers by different authors are included. The main topic dealt with is the guidance and control of missiles. Specific papers consider the range of problems from the selection of a missile guidance and control system to its final testing in actual flight.</p>	
<p>North Atlantic Treaty Organization Advisory Group for Aeronautical Research and Development</p> <p>SECOND GUIDED MISSILES SEMINAR GUIDANCE AND CONTROL 24-28 September 1956, Venice, Italy</p> <p>Twenty-four papers by different authors are included. The main topic dealt with is the guidance and control of missiles. Specific papers consider the range of problems from the selection of a missile guidance and control system to its final testing in actual flight.</p>		<p>North Atlantic Treaty Organization Advisory Group for Aeronautical Research and Development</p> <p>SECOND GUIDED MISSILES SEMINAR GUIDANCE AND CONTROL 24-28 September 1956, Venice, Italy</p> <p>Twenty-four papers by different authors are included. The main topic dealt with is the guidance and control of missiles. Specific papers consider the range of problems from the selection of a missile guidance and control system to its final testing in actual flight.</p>	



NORTH ATLANTIC TREATY ORGANIZATION
ADVISORY GROUP FOR AERONAUTICAL RESEARCH AND DEVELOPMENT
(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

A G A R D
SECOND GUIDED MISSILES SEMINAR
GUIDANCE AND CONTROL

0001093 8



Papers presented at an AGARD Guided
Missiles Seminar, held 24-28 September 1956,
in Venice, Italy

The papers included in this AGARDograph were obtained through the courtesy of the United States Air Force, the United States Army, the United States Navy, and the Aircraft Industries Association. Colonel C. E. Evans of the ARDC organized a committee consisting of himself, Lieutenant Colonel Roland V. Tiede, Commander William M. Romberger, and Major General J. F. Phillips, who obtained the papers from their respective organizations, and made arrangements for the presentation in Venice.

This publication supported by the United States Air Force under Contract No. AF 18(603) - 133, monitored by the Air Force Office of Scientific Research of the Air Research and Development Command.

INTRODUCTION

The papers presented in this volume survey one of the major divisions of guided missiles design, and provide information which will be useful to those engaged in the design of missiles and missile systems.

In addition, the Seminar illustrated the value of discussing matters pertaining to guided missiles in international meetings and pointed the way towards future activities in this field.

THEODORE VON KÁRMÁN
Chairman, AGARD

TABLE OF CONTENTS

	Page
Introduction	i
1. Weapons System Philosophy, George H. Clement	1
2. ORO Weapon System Philosophy, Harlan C. Meal	11
3. New Principles in the Design of Superior Communications, Navigation, and Missile Guidance Systems, W. P. Lear, Sr.	29
4. Guidance Techniques, Walter L. Webster, Jr.	39
5. Considerations in the Choice of a Missile Guidance and Control System, Robert W. Mayer	57
6. Inertial Navigation, Norman F. Parker and Charles P. Greening	67
7. Aiding the Inertial Navigation System, William F. Ballhaus and Frederick Stevens, Jr.	87
8. Linear Homing Navigation, Robert K. Roney	101
9. Pitfalls in Missile Control, Robert L. Johnson	115
10. The Effects of Airframe Characteristics on Control System Design, F. E. Perry	139
11. Geometrical Stabilization Based on Servodriven Gimbals and Integrating Gyro Units, Charles S. Draper and Roger B. Woodbury	153
12. Sampled-Data Systems, John R. Ragazzini	183
13. Digital Techniques in Missile Guidance Systems, Sidney Darlington	229
14. The Use of Digital Computer Techniques in Missile Design and Control, D. H. Gridley	249
15. The Application of Noise and Filter Theories to Guidance Problems, R. J. Parks and Robert M. Stewart	265

TABLE OF CONTENTS
(Continued)

	Page
16. Recent Developments in Fixed and Adaptive Filtering, A. G. Carlton and J. W. Follin, Jr.	285
17. Practical Problems Encountered in Missile Guidance and Control Design, R. E. Whiffen	301
18. Application of Methods of Science to the Problem of Reliability, C. Raymond Knight	315
19. Reliability of Guided Missiles, Edwin A. Speakman	331
20. Laboratory vs. Flight Evaluation of Airborne Guidance Components, W. H. Clohessy	341
21. Trends in Field Testing of Guided Missiles, Ernst A. Steinhoff	347
22. Low Signal Level Missile Instrumentation, L. G. deBey	359
23. On the Way to Automated Processing of Flight Measurement, W. E. Klemperer	375
24. Paper on the Guidance and Control of Missiles, Stephen Waldron	387

WEAPONS SYSTEM PHILOSOPHY

George H. Clement*

SUMMARY

In achieving an effective defense and an increased weapons capability, there has been an ever larger expenditure of our national resources. The difficult problem exists in maintaining an adequate level of defense with a minimal drain on national resources. The weapons system philosophy is essentially a point of view embracing a technique for ordering, classifying, and analyzing a technologically complex mechanism, organization or process so that each element and each problem can be considered in its proper context and treated in its proper perspective. The objective of such an approach is to attain the objectives of the system as a whole with a minimum expenditure of resources. Intelligently applied, this approach can be a powerful tool aiding the design, development, and management of military weapons systems.

SOMMAIRE

Dans la réalisation d'une défense efficace et d'une capacité accrue des armes, il y a eu toujours une dépense plus grande de nos ressources nationales. Il existe un problème difficile, celui de maintenir un niveau de défense adéquat en dépensant au minimum les ressources nationales. La philosophie du système d'armes est essentiellement un point de vue embrassant une technique qui vise à ordonner, classer, et analyser un mécanisme dont la technologie est complexe, une organisation ou un procédé, de telle manière que chaque élément et que chaque problème puisse être considéré dans son propre domaine et traité dans sa propre perspective. L'objet d'une telle approche est d'atteindre les buts du système considéré dans son ensemble avec une dépense minimum des ressources.

1. INTRODUCTION

Men have long been concerned with the defense of their national resources and their national way of life against the possible encroachment of a potential attacker. Military forces have been continually designed, developed, organized, and then modified to counter the increasing threat imposed by technological advance and by the changing social structure of the world in which we live. The function of these military forces has been to deter any potential attacker from

carrying out an aggressive action by increasing the probability of his failure.

In the realm of modern military affairs we continue to desire an ever increasing capability to satisfy our defense needs, and no one, I believe, would argue against the fundamental wisdom of seeking this increased capability. The margin of superiority in the performance of our weapons, for example, the range, speed, altitude, and payload capabilities of our airplanes compared with those of a potential enemy may spell the difference

* The RAND Corporation, Santa Monica, California.

between victory and defeat in a possible future conflict. What we often fail to appreciate is that the increased capability we need and strive to achieve is purchased at the cost of an increased complexity of the equipment and organization that results.

The consequence of this increased complexity is to require an ever larger expenditure of our national resources to achieve adequate levels of defense effectiveness. The resources that we have at our disposal are always limited and have very real alternative uses. Fissionable material expended in the construction of bombs cannot be used for the generation of electric power. Manpower committed to the maintenance of the military forces cannot be used in any productive capacity in the national economy. A difficult problem exists in achieving and maintaining an adequate level of defense effectiveness with a minimal drain on our national resources.

The modern implements of warfare are an order of magnitude more complex and more expensive than their predecessors of a few years ago. The increased range, greater speed, and higher altitude of our bombers has resulted in the need for better navigational aids, bombing and fire control mechanisms, and communication equipment; also longer runways, more highly trained aircrews, and more extensive maintenance and repair organizations. The airplane as a weapon has become more than a simple assemblage of powerplant, fuel, airframe and controls. It is now a complex and highly integrated mechanism consisting of ground and airborne components which we call a weapon system. I would like to spend the next few minutes exploring the potential of a particular point of view, which we call the weapons system philosophy, as it applies to the development of military equipment.

2. THE SYSTEMS CONCEPT

When we speak of weapons we usually think of the individual elements, the tools of warfare. Firepower on target is thought of in terms of bullets, bombs, and grenades. Mechanisms for transporting this firepower to the vicinity of the target brings to mind such things as guns, tanks, trucks, airplanes, and guided missiles. The people who operate the equipment are thought of as infantrymen, pilots, navigators, radio operators, and gunners. The devices of communication are thought of as telephones, radios, and radar sets. In a sense each of these things taken by itself is a weapon. But I think it is readily evident that each of these taken by itself is valueless. A bullet is of little military use without a gun, without a soldier to fire the gun, without some means of communication to direct the effort of the soldier, and so on.

Is then a weapons system an aggregation of these components? In a sense yes, but it implies more than this. The weapons system philosophy is intimately concerned not only with the components of warfare singly and in aggregate but also with the interactions and interrelationships among these components, with the expenditure of resources necessary to bring the military force into being and keep it operating at an adequate level of effectiveness, and with the process of use of the integrated whole as a means of achieving some desired military purpose.

What then would be an adequate definition of a weapons system? I think that we have to recognize that the concept of a weapons system can mean different things to different individuals, depending upon their point of view. The particular concept of a weapons system can represent different levels of aggregation depending upon the interest and concern of the individual viewer. Thus the head of a national government may be concerned with a social system, a political

system, an economic system, and a military system as the instruments of national policy. The man in charge of the military system views it in turn as being composed of an air force, an army, and a navy; in other words, an air, a ground, and a sea system. The chief of the air force thinks in terms of an air defense system, a strategic air system, a tactical air system, an air transport system, and so on.

Skipping a few steps, one can view the airplane itself, the interceptor in the air defense arm, as a system. The interceptor in turn contains such things as a propulsion system, a navigation system, a communication system, a fire control system, and so on.

For the purposes of this discussion, I think we can limit ourselves to considering such things as the air defense system, the strategic air system, or the tactical air system as being weapons systems at the highest level of aggregation with which we will have to deal. Accepting this concept, the complete weapons system includes then not only the vehicle together with its armament, communications, navigation, fire control, and other systems; but also, and as an integral part of the system, the supporting facilities, the bases, the communications net, the maintenance and test equipment, the logistical supply and transport, the training installations, the mechanics, and other specialized personnel. All of these components are just as much a part of the weapons system as the vehicle itself.

It is useful when thinking of weapons systems and their components to regard them as being different orders of the same kind of thing, all members of a continuum of systems, supersystems, and subsystems. To the propulsion engineer an engine is a system. At the same time it is a member of a higher order system, the aircraft in which it is installed. Further, the engine is composed

of many lower order systems or subsystems such as an ignition system, a fuel system, an exhaust system, and still others. We see then that the systems approach, the weapons systems philosophy, is in part a technique of classifying and ordering a complex mechanism or process so that each element or problem can be considered in its proper perspective.

The systems approach is not a new concept nor did it find its origin in dealing with military problems. Systems analysis, systems engineering, the systems concept has been applied for many years by the engineers and managers of such industrial enterprises as large telephone companies and the producers and distributors of electric power. As we have seen, the systems approach arises out of complexity and represents a method of ordering and controlling for effective use a technologically complex structure.

3. A TYPICAL AIR DEFENSE SYSTEM

At this point it will be useful to examine a typical weapons system such as an air defense system. As we discussed earlier, the purpose of a defense system is to deter any potential attacker from carrying out an aggressive action by increasing the probability of his failure over that which he might expect in the absence of our defensive effort. In fact we seek to maximize the probability of his failure by constructing as effective a defense system as our limited resources will permit.

A convenient measure of the effectiveness of an air defense system is a quantity we can call the "kill potential." Kill potential is defined as the expected number of bombers destroyed by the defense system under some standard condition. In choosing between two air defense systems that can be developed and operated for an equivalent expenditure of

resources, the system having the higher kill potential would be preferred. Similarly, actions which tend to increase the kill potential of the system as a whole would be considered desirable, while those that tend to reduce the kill potential would be considered undesirable. Now the operational functions of a deployed air defense force are concerned with the detection, identification, interception, and destruction of hostile aircraft.

In order to detect hostile aircraft, some method of surveillance of the air traffic in the region to be defended must be provided. This is commonly accomplished by means of a radar surveillance net deployed over the geographic area to be defended, and further, in order to maximize warning time, reaching as far beyond the perimeter of the defended area as is practical. These radar installations may be located at fixed sites on land and augmented by picket ships and picket aircraft along coastal frontiers. The radar net may further be supplemented by acoustical and optical detection nets such as operated in the United States by the Ground Observer Corps. This detection net is responsible for acquiring raw data pertaining to the current air traffic situation, mapping individual aircraft or groups of aircraft, and determining their altitude and their number. The individual detection installations must be tied together by means of an extensive communication system over which the raw surveillance data can be transmitted to identification facilities.

At the identification facility, the raw surveillance data are received, processed, and evaluated. Individual pieces of air movements information are correlated to establish aircraft tracks. In this process, supplementary data may be requested and received from the detection sites. After the tracks are established, they must be identified as friendly or hostile. Unknowns must be

subjected to sufficient investigation to determine their status. The individual tracks are summarized to describe the air situation in the geographic area for which the identification facility is responsible. This air situation information then must be communicated to identification facilities responsible for adjacent geographic areas and to the appropriate interception facilities.

At the interception facility the air situation data are received and the threat is evaluated. On the basis of this evaluation an appropriate fraction of the available defensive force is committed to the air battle. Orders are issued to scramble interceptors and to bring surface-to-air guided missile and antiaircraft artillery units into action. Information on the position of the hostile aircraft is communicated in the form of guidance instructions to the appropriate weapons units of the air defense force.

At the weapons units, the guidance information received from the interception facility is employed to bring fire power to bear on the hostile aircraft. Interceptors and surface-to-air guided missiles are vectored to the vicinity of the hostile aircraft where their airborne search, detection, and tracking equipment enables them to close on the enemy and bring the hostile aircraft under fire. Antiaircraft artillery units receive position information on hostile aircraft so that their fire power may be properly employed.

Technological advances in the aircraft art are continually increasing the range, altitude, and speed of bombardment vehicles. These performance increases, particularly the increase in speed, means that over the years, less and less time will be available to bring the air defense force into action. To counter the continually shrinking time available between initial detection and the bomb release time, development of equipment to perform

many of the detection, identification, interception and destruction functions of the air defense force by automatic or semiautomatic means is under way. Since the detection, identification, and interception facilities of an air defense force are commonly referred to as the "ground environment" this new equipment is being called the Semiautomatic Ground Environment or SAGE System.

At this point one might well say, "How does this affect me?" This is all well and good but how does this apply to problems that arise in the development of missile guidance and control systems? How does this affect me? To illustrate this let us for the moment assume that the guidance and control system we are concerned with is to become a part of a surface-to-air missile which in turn is needed as a component of an air defense system such as the one we have just described.

4. THE DEVELOPMENT OF A SYSTEM

As we all recognize, the process of design and development is one of compromise. Decisions have to be made throughout this process among alternative methods of achieving particular goals. For example, should the missile guidance system contain some means of homing on the target? If so, should the missile carry an active radar for terminal guidance or should it make use of energy reflected from the target aircraft when illuminated by a ground installation? Or should the ground installation track the present position and velocity of both the target aircraft and the missile and on the basis of this information compute the proper maneuver to be performed by the missile and issue instructions to it to do so?

The resolution of these questions cannot properly be made on the basis of examining the guidance and control function in isolation, nor even on the basis of examining the

performance of the missile by itself; rather it must include a consideration of the influence of such questions on the operation of the air defense system as a whole.

Some might argue that we are concerned here with the accuracy of the missile; that it is obvious that if the missile can be made more accurate then it will certainly result in more kills.

It is not at all clear, however, that making the missile as "accurate as possible" will necessarily result in maximizing the effectiveness or kill potential of the defense system. In this case, the numerical value of the kill potential is given as the product of three quantities. First, the number of missiles that can be launched during the engagement; second, the probability that a given missile will not malfunction or result in a gross error; and third, the probability that a non-malfunctioning missile kills the bomber. In our example an increase in the accuracy of the missile would result in increasing the value of the third factor in the kill potential equation, the probability that a non-malfunctioning missile would kill the bomber. This would result in an increased kill potential only if the value of the first two factors in the equation are not materially degraded.

If in the course of achieving the greater accuracy, additional guidance and control equipment were added to the missile, the resulting increase in weight might reduce the range or speed of the missile which in turn could result in reducing the number of missiles that could be launched during the engagement. Further, the added equipment might adversely affect the reliability of the missile. Either or both of these possibilities then might well reduce the effectiveness of the system as a whole in spite of the enhanced accuracy of the unit. I think it is clear, then, that the manifold decisions and choices among alternatives that must be made during the

course of the design and development of a missile must include a consideration of the influence and feedback on the weapons system as a whole as well as the effect on the particular component in question.

From the point of view of the missile designer, then, it is important to have an appreciation, an understanding of the military environment in which the device will operate. It is necessary to appreciate the relevant technical, operational, and logistical factors of this environment to understand the interactions and interrelationships among these factors and how they influence the operation of the system as a whole. I like to think of these technical, operational, and logistical factors as being the anatomy, that is, the structure of the weapons system. The technical factors include such areas as propulsion, fuels, guidance, control, communication, and the like.

Engineers are accustomed to think of the performance of military air vehicles in terms of such quantities as range, speed, altitude, rate of climb, maneuverability, and accuracy. It is generally held that an excellent vehicle is one in which the values of these performance parameters are made as large as possible. The level of performance actually achieved with a particular vehicle, however, depends not only upon the current level of technology, but also upon the design choices that are made in the course of the development of the vehicle.

I think we all recognize from long experience that the design and development choices we make are likely to affect all, not just one of the performance parameters. For example, the choice of the powerplant can influence not only the speed of the missile but also its range, altitude, rate of climb, and maneuverability as well. It follows then that quite independent from any consideration of how the desired performance parameters

were specified, the final result will, in all likelihood, represent a compromise in which many of the desired performance levels are only partially satisfied. The question is then "What are good compromises?" To answer this question we must examine the rest of the weapons system anatomy, the operational and logistical factors.

The operational factors which apply include such considerations as the mobility of the system, its state of readiness, its reliability, its susceptibility to countermeasures, and its data requirements. The specific meaning attached to these considerations and their influence on the effectiveness of the endeavor depends upon the character of the weapons system with which we are dealing, that is, whether it be an air defense system, a strategic air system, or a tactical air system.

Mobility of the system enables the force to be rapidly deployed and moved as the occasion demands. It enables the concentration of firepower at times and places where it is needed. As applied to missiles, the operational concept of mobility means that the effective range and speed of the system should be as great as possible. This range and speed can be supplied either by the inherent range and speed of the missile or by the capability of the weapon system for rapid deployment and movement to new bases of operation.

The state of readiness of the system is its ability to respond quickly when called upon. In an air defense system only those missiles that are available can be of any use when the attack comes. If the guidance or control equipment in the missile requires an excessive warmup or preparation time, the missile will be of little use to the defense operation.

Reliability refers to the probability that the system will function within its design limits whenever it is called upon to perform. Again, for example, if the guidance and control system is mechanized by equipment of such a delicate nature that it cannot be maintained under normal operating conditions; or if in an effort to squeeze out the last bit of performance the equipment is pushed into operation at levels beyond its capacity; or if the equipment is continually subjected to unknown or poorly understood environmental conditions, a high failure rate will result. The gross errors, malfunctions, and excessive maintenance that follow do not contribute to the effectiveness of the defense system.

The data requirements of the system have to do with the quantity, character, and rate of flow of the information that is essential to the operation. Before a defense system can bring firepower to bear on an attacker, the bombers must be detected, tracked, and identified. Again, if the data requirements of any component, such as the guidance and control equipment, are excessive or unusual or if they increase the susceptibility of the system to countermeasures, then the performance of the system as a whole will suffer.

Turning now to the logistical factors we include consideration of such items as facilities, equipment, supply, maintenance, personnel, and training. Logistics provides the means for the conduct of air operations and is an important factor in determining the effectiveness of the weapons system. During the course of the development of our surface-to-air missile the designer, either explicitly or implicitly, is determining the logistical requirements of the system. Before deciding to use that unusual compound as a fuel, on the basis of, say enhancing the range performance of the missile, the logistical

implications of the decision should be investigated. Can it be manufactured and transported to the point of use in the required quantities? Among its properties is the proposed fuel so toxic and corrosive that it requires inordinately complex supply, transport, storage, and handling procedures? If so, the overall performance of the system may be degraded in spite of the enhanced range that would result.

Now that we have examined the anatomy of a weapon system and found that it contains many elements which place seemingly conflicting requirements on the system, I think it is fair to ask again "How can one determine what is a good compromise?" How can one determine a preferred course of action in the midst of this welter of conflicting detail? To investigate this question let us again restate our understanding of the weapons system philosophy and then examine the role of operations research as applied to the problem of weapons system development.

The weapons system philosophy is a point of view, embracing a technique for ordering and classifying a technologically complex mechanism, organization, or process so that each element or problem can be considered in its proper context to the end that the objective of the system as a whole may be realized with a minimal expenditure of resources.

As we noted earlier, the system approach is not a new concept but has been applied for many years as an aid in the design, development, and management of large and complex industrial enterprises. The first formal application of this technique to military problems appears to have been made by the British during World War II, when they developed and employed a management device that has come to be known as operations analysis. At that time the British were concerned with determining how best to

employ the weapons, equipment, and personnel they then had available to most effectively accomplish an assigned military mission. The success of this technique in conserving resources or in maximizing the military effectiveness of a given force was soon apparent. Other nations quickly recognized the benefits of this technique, and applied it to their own problems of how to obtain the maximum effectiveness from existing weapons systems. As we noted earlier, the trend toward greater technological complexity of military weapons systems has continued and is continuing at an ever accelerating pace. I think it natural, then, that the systems analysis technique be applied not only to the use of existing weapons systems but should be extended and applied to solving problems that arise during the course of the design and development of weapon systems as well.

The application of the systems analysis technique, or operations research as it is sometimes called, to problems arising in the design and development of weapons systems affords us the opportunity of arriving at decisions that are in the proper context, that are consistent with the overall objectives of the weapons system. It provides the framework for forming judgments and helps us to phrase and select the criteria on the basis of which we can determine an answer to our earlier question "What is a good compromise?"

5. PRINCIPLES FOR ANALYSIS OF WEAPONS SYSTEMS

Dr. Charles Hitch, of the RAND Corporation, has developed principles that are of such importance in analyzing weapons system development problems that I would like to repeat two of them here. The first point I wish to draw from Dr. Hitch's work is that in analyzing a weapons system problem, the

effects of a proposed action must be assessed at least at the next higher level of aggregation, and further, must be assessed in terms of the operation or function that the weapons system is performing.

Recalling our earlier example when we were discussing the influence of attempting to increase the accuracy of the surface-to-air missile guidance and control system, we learned that an increase in missile accuracy would not necessarily be consistent with the objectives of the weapons system at its next higher level of aggregation. That is, an increase in missile accuracy would not necessarily lead to an increase in the kill potential of the air defense system; in fact, it might even materially reduce the overall effectiveness of the system. It was not good enough to assess the influence of the proposed action on the basis of its effect on the performance of the missile. The effects of the proposed action must be assessed at least at the next higher level of aggregation, and must be assessed in terms of the operation or function the weapon system is performing.

The second point that I wish to draw from Dr. Hitch's work also concerns the selection of criteria in terms of which the consequences of some proposed course of action are assessed. In analyzing the influence of proposed development alternatives on the effectiveness of a weapons system, it is a common and often fatal mistake to select criteria that concentrate attention on a single input; to maximize some objective's function, say bombers killed, for a given quantity of input, say number of picket aircraft in the detection net, or to minimize the number of picket aircraft required to achieve a certain number of bombers killed.

To do this completely overlooks the resource substitution possibilities that may exist. An objective can usually be achieved by various combinations of resources. An air

defense system of any specified level of effectiveness may be achieved with very different mixes of ground environment, interceptors, and surface-to-air guided missiles. Some resource substitution possibilities are direct and obvious; many others are indirect and subtle. Judicious application of the operations research technique to weapons system development problems will uncover the fact that the possibilities of substitution in carrying out any operation are typically much greater than are at first apparent.

At this point I would like to interject a note of caution. The weapons system philosophy like any useful tool can not only be properly and usefully applied, but it can also be wantonly misapplied with devastating results. The weapons system philosophy and its related concepts and techniques, systems analysis or operations research, are powerful tools. Properly used they furnish a basis, a framework, that helps us think through and understand a complex problem. It represents in no way, however, a substitute for thought. If the techniques we have been discussing are blindly, unimaginatively, or unthinkingly applied only rigor mortis can result.

Proper application of the weapons system approach to design and development problems enables us to choose from among a set of alternatives those that are good as opposed to those that are poor. It may even permit us to rank the choices according to some scheme of preference from best to worst. However, searches for the "optimum" or "best possible" alternatives are likely to be not only time-consuming, but also fruitless. The fundamental reason for this, I believe, is that the analytical treatment of any problem requires the construction of an analytic model which approximates the real world situation. The degree of approximation can range from quite crude to very sophisticated, but it still remains an approximation.

While very useful inferences can be made from a study of the model, it should be kept in mind that it is a model, an idealized approximation, and one should guard against over-analyzing or using the model to obtain results in areas for which it is unsuited.

A second possible misapplication of the weapons system philosophy can arise in the field of basic research. We have been discussing today the weapons system philosophy and how it applies to the design and development of military equipment. I think we have to recognize here that there is a characteristic difference between development and fundamental research. Development activities are aimed at a specific goal, a new or improved weapon, machine, or process. Basic research has no such specific goal, it does, however, have a purpose, understanding nature in all its bewildering and infinite variations. Fundamental science searching through this wilderness adds piece by piece to our store of knowledge. As each new fact is brought to light something that was unknown or uncertain becomes known, becomes measured, determined, and verified. I think it is also universally appreciated that in this technological age a strong military, in fact a strong society, must rest on a firm base of fundamental science. In fact, all technology has its roots in and grows out of basic scientific understanding. It is in this fact and in the fact that basic research is essentially an exploration of the unknown that I think the danger lies. In most applications of the weapons system philosophy we are interested in determining a preferred course of action from an analysis of the predictable consequences that would result if any of a number of alternative actions were taken. How can one predict the consequences of the unknown? It must first be found, discovered, measured, and verified before it can be evaluated. If the pattern of thinking represented by the weapons system philosophy were to be misemployed, say in an effort to

determine in advance the worth of a proposed fundamental research program, the result of the analysis would not only be meaningless, but such a pattern, if allowed to become established, would strike at the very foundation of our strength.

I would like to extend this argument a little further, a short way into the realm of what some might consider to be applied research or development. Often in the course of the development of military equipment and usually at the component level we venture into the unknown or at least into the realm of the uncertain. We desire to put together in a single machine several functions, several technological processes, that have never before been related. The mutual compatibility of the functions working together in a common environment is quite uncertain. In the course of developing such a device it is often suggested and often profitable to try separate and independent approaches. In this case, as in the case of basic scientific research, it is my belief that trying to

predict in advance on the basis of an analysis which approach will be successful, which one will work and which one will not, or which one can be developed to an acceptable level of performance in the shortest time, is not only a fruitless activity but a dangerous one. Why dangerous? Simply because the problem contains a fundamental uncertainty which no analysis can remove and one might be led to follow the wrong course.

In conclusion I would like to restate that the weapons system philosophy is essentially a point of view, embracing a technique for ordering, classifying, and analyzing a technologically complex mechanism, organization, or process so that each element, each problem, can be considered in its proper context and treated in its proper perspective to the end that the objectives of the system as a whole may be realized with a minimal expenditure of resources. Intelligently and thoughtfully applied, it can be a powerful tool aiding in the design, development, and management of military weapons systems.

REFERENCES

1. Hitch, Charles, "Sub-Optimization in Operations Problems," Journal of the Operations Research Society of America, Vol. 1, No. 3, May, 1953.

ORO WEAPON SYSTEM PHILOSOPHY

Harlan C. Meal*

SUMMARY

The function of operations research is to select and organize information which will assist military planners to arrive at decisions leading to an orderly succession of weapons systems. The task is complicated by the rapid technological advances which often render a weapon obsolete before it is even placed in production. All the factors such as study, research and development, construction of prototypes, testing, preproduction, production, training, development of tactical doctrine, and finally deployment and operation of weapons must be taken into account before settling on the adoption or discontinuance of a program. Selection of weapon program alternatives must be made with the effectiveness of the whole defense system in mind. A sound weapons program must provide continuing effectiveness over a long time. For this reason, a continuing renewal of the analysis of weapons systems, and their effectiveness in the presence of changing threats is a task that is undertaken by those engaged in operations research.

SOMMAIRE

Le but de la recherche en opération est de sélectionner et d'organiser l'information qui assistera les projeteurs militaires, à arriver à des décisions concernant une succession ordonnée de systèmes armés. La tâche est compliquée par les progrès technologiques rapides qui souvent rendent une arme désuète avant même qu'elle ne soit mise en production. Tous les facteurs tels que l'étude, la recherche et le développement, la construction de prototypes, les essais, la preproduction et la production, la mise en route, le développement d'une doctrine de tactique et finalement le déploiement et l'opération des armées doivent être pris en compte avant la mise sur pied et l'adoption ou le rejet d'un programme. Un programme d'armes solidement établi doit pouvoir procurer durant une longue période, une efficacité continue. Pour cela, une analyse continuellement renouvelée des systèmes d'armes et de leur efficacité en présence de menaces changeantes a été entreprise par les personnes travaillant dans la recherche en opérations.

The goal of present Army programs is to provide a high level of effectiveness over a long period of time. The object of these programs is to deter, therefore there is no peak year, no race to build a particular weapon system to maximum effectiveness at any particular time. These considerations, which apply to guided weapon systems as they do to all Army programs, can be said to embody the

current Army weapon system philosophy. I will not elaborate further on this philosophy but instead illustrate how that philosophy is practiced. Specifically, I want to show how the Operations Research Office of The Johns Hopkins University is attempting to develop a technique for supplying the Army with better information on which to base its planning decisions.

*Operations Research Office, The Johns Hopkins University.

A continuing high level of effectiveness is achieved by planning research and development and procurement programs which lead to an orderly succession of weapons systems. Technological advances occur at such a rate that, in order to carry out this planning task, a great deal of information is required. This must be selected and ordered information, not just all the information. The process of selecting and ordering this information will illustrate how the Army philosophy is practiced, since the information used in any planning process will influence, if not determine, the resulting plan.

In the analysis of weapons systems, either offensive or defensive, we start with a given capability of a weapon. This comes from the weapon designer. As a result of a request by the military, usually in the form of military characteristics, the weapon designers produce a design proposal for a weapon which will shoot a certain distance, with a specified accuracy, carrying a listed payload, and so forth. Although a great deal of improvement has occurred, it is often in the specifying of military characteristics that planning difficulty first occurs. Often the military planners are all too familiar with the tactical problem and the kind of weapon which would solve it, if it could be built, but are not familiar enough with the state of the weapon art to ask for a feasible weapon or weapon system. For example, a set of military characteristics might call for an antiaircraft missile which has a range of 20 miles, an altitude capability from zero to 60,000 feet, probability of kill against a single aircraft (either alone or in formation) of 0.8, a missile weight of not more than 600 lbs., and with the entire unit having the mobility of a standard military truck.

I am sure that the absurdity of such a combination of characteristics is obvious, but it may not be so obvious to the military man who has been many times amazed by the

accomplishments of the research and development organizations, and has gradually been convinced that they can do anything. Operations research can assist both the military planner and the weapon designer by translating capabilities into effectiveness and by translating tactical effectiveness into requirements to develop an increased capability.

This is just one example of how difficulties in selecting and ordering information occur. There are many such examples. To get to the research and development groups the information which they need to design good weapons, to keep the military organizations informed as to the current state of the weapon art, to relate both of these to the attack capability and the cost of proposed programs, and to do this in military rather than technical terms, requires much handling of information. The way in which this is done illustrates how the philosophy is put into practice.

Let us now look at how this process actually takes place. Starting initially with proposed new weapons coming up to fill a need, how does the Army decide which one, or what mixture, to adopt? The operations research teams examine the problem in this fashion. Referring to Fig. 1, in taking an initial estimate of the problem, the operations analysis is certain to be very unsophisticated. The weapon designers estimate that they can have the weapons by 1958 which will have certain capabilities. Intelligence estimates that in 1958 the air threat is, say, 8-1200 bombers, similar to U. S. B-50's, flying at medium altitude. The effectiveness study combines these two with some simple mathematics and estimates that it will take 500 units of the type in question to limit the damage the enemy bombers can do to the defended target system to some specified amount. Furthermore, these units will have a certain cost.

In doing this first analysis of the problem, a check is made that the design proposed is technically feasible and that reasonable lead times for development, testing, tooling, training, production, and deployment have been allowed for. No very detailed study of tactics is made, but at least a part of the total spectrum of enemy courses of action will be examined to see what the enemy can do that hurts most. Most common among such simplifications is that the effect of counter-measures will be ignored and that low altitude attacks need not be considered.

The analysis of effectiveness then becomes quite simple, or at least so it seems after the fact. Commonly, assumptions are made to avoid considering the more complex of the operational questions. An example is the statement that fire will be distributed either uniformly or at random over the attackers in any particular engagement. A more sweeping kind of assumption is that one can look at only part of the weapon system, for example, the weapon itself, ignoring the information gathering and control system.

The problem of cost is not very sophisticated either. It may simply state that the cost of the required number of weapons is a billion dollars and it will cost a hundred million dollars a year to operate them. Comparison of the results of such analyses for several systems allows the comparison of cost-effectiveness ratios, the weapon giving the minimum cost over effectiveness being the best.

Actually, I have already skipped the real first, or most elementary analysis which develops a figure of merit for a unit and the cost of that unit. For example, it has been said in the past that the kill potential or expected number of kills per battalion, divided by the effective annual cost, gave an effectiveness-cost ratio which could be used to compare weapons. Effective annual cost is the

annual cost of the system, plus an amortization of the initial investment over the expected useful life of the system, usually four years.

What is wrong with such an analysis? Fundamentally it ignores the effect of time and that things change with time. Since it is the case that no tactical or strategic solution will remain pertinent for long, it is also the case that analysis of a weapon or weapons system program, even if directed toward a specific time, must reflect these changes.

In the past, policies were worked out slowly over a period of years and modified slowly. A weapon would be designed, a prototype constructed and extensively tested, a few models fabricated and put into the hands of the military user to train with and establish doctrine, long before any thought or question of quantity production came up. In fact, usually no significant production would take place until hostilities were imminent or had already broken out. The time scale was so long that it could be taken to be infinite, that is, the effects of changes with time ignored.

The time scale now is not much shorter in years than it was twenty years ago. The difference is that we move through it much faster. Weapons and weapons systems are designed, built, tested, and produced in quantity, and discarded as obsolete, without ever being used in combat. In addition, the technology moves so fast that even before a weapon is in production the ideas which make it obsolescent exist; before the troops have become familiar with a new weapon, the superseding weapon is well into development. It must be so, since offense and defense both have rapidly moving technologies which they both must exploit if they are not to fall behind. It is a little like the legendary Red Queen who had to run as fast as she could in order to stay where she was.

In order to run even a little faster, an organizational or planning scheme, a philosophy, must exist which allows, in fact requires, the effects of time changes to be shown in the analysis of operational value of any weapon or succession of weapons. In a rapidly changing technology a static analysis, which does not show the effects of changes with time, is worse than valueless, it is dangerous.

Let us look now at how the analysis scheme shown in the first chart must be modified in order to supply planners with information relative to the effectiveness of a predicted defensive capability in meeting a predicted offensive capability. More importantly, the scheme must be able to show how a succession of offensive capabilities can be met by an orderly development and deployment of more and more advanced weapons.

First, a detailed projection of the offensive capability is needed. This can be obtained to a certain extent from intelligence sources, but it is also based on projections from the known state of the art. That is, if the enemy has knowledge of certain kinds of microwave power sources, then it is certain that he can have a certain kind of radar, after a lead time which is predictable within fairly narrow limits. In this way, the availability of a wide variety of offensive capabilities can be predicted. Note that here it is not enough to predict a succession of weapons of the same type as, for example, the replacement of a piston-engined bomber force by a jet bomber force. Also, new weapons systems and the capability of using new tactics must be included. The phasing of different types of electronic countermeasures is an example. The arrival of ballistic and cruise missiles is another obvious kind of threat which must be included. But, to repeat, it is not enough simply to include it. For any particular

offensive capability, the same availability analysis as for the defensive capabilities must be done. It should specify, insofar as possible, ranges, accuracies, etc., and the dates these can be expected to be operationally available.

In a similar fashion, the spectrum of defensive capabilities is predicted. This is easier, however, for we can go and confer with the designer and determine what difficulties he expects to encounter in the process of developing the device in question. Also, and very important, we can find out how he proposes to modify the present design with developments expected to materialize from ideas already existent. A good example of this is the use of transistors. Given a guided missile which is constructed using vacuum tubes, when will transistors be used and what improvements can be expected in reliability, weight reduction, lower noise figure receivers, and other factors.

It is not enough to confer only with the designer, but the engineering and production staffs must estimate how long it will take to get the modification into production, produce it, and modify the existing equipment. Only if this is done in a detailed and painstaking manner, however, can the analysis produce results which are of positive value to the planners.

Fig. 2 shows what the result of such a study might be. This is the same type of lead time analysis as is done in the static case mentioned before. This is, however, only the beginning, for at some point in the process a new idea, an invention, may appear and we desire to know what effect, if any, this can have on the program. For example, advances in radar technology may make desirable the addition of a seeker to the missile, if it is an air defense missile, or if it is an offensive weapon program, a new navigational technique may appear to be feasible.

If it is very early in the program, it may be that the original idea will be modified and all changes can be accomplished during the research and development phase. Later in the program it may be decided not to incorporate the modification until the first program is complete and then apply modification on units already deployed. All weapon successions appear as shown in Fig. 3, but here the time scale is somewhat compressed. The study phase might occur very early and all the phases be accomplished at such a rate as to catch up by the time production is started. In such a case the phases of the initial program will be somewhat longer in order that the modification can be completely engineered into the system, but the original design will have to be modified. Early testing will probably be with the original model, but will be somewhat extended to accomplish testing on the modified version. I am sure that all of this is quite familiar, but the effect that it has on timing the program in predicting when capability will be available, is not so often clear to the military planners, and I might add, the amount of delay is often underestimated by the weapon designers.

Let us look quickly now at a slightly different case, a case of the kind which is much more important in predicting what effectiveness over time will be. In this situation Fig. 4 shows the new idea for improved capability somewhat later in the cycle. It is not feasible to incorporate the new modification in the weapon prior to production. It may delay the system or it may be of such a nature that it cannot be incorporated into the older version. One should note especially how the phases overlap. This may be a parallel weapon development where the weapons may be designed with similar capabilities, but with the second at lower cost. In any case, planners and executives are faced with overlapping programs and may be forced to decide at any time whether one or the other is to be cancelled or accelerated.

It may be that the earlier availability of one system is a requirement to fill a need; at the same time it may be so much less efficient that it should be terminated in favor of the later one after a very short operational life. Constructing of lead time charts such as these is a tedious process and requires a good deal of effort, but without them it is not possible to determine how effectiveness can be maintained, or increased, over time.

Nearly the same kind of analysis of availability must be made for the offensive capabilities. This would result in a time-phased threat picture as shown in Fig. 5, which would show how different weapon systems are phased into and out of the total offensive weapon system. Piston-engined bombers are discontinued with the production of more advanced bomber types. We can predict how soon given numbers of these new bombers can be available. Also, from our knowledge of offensive and defensive technologies, we can predict when the attacker can attack with missiles launched from submarines. These might be either ballistic or cruise missiles, and so on, for the other threats shown in the figure.

The figure is very much simplified. It does not show how the countermeasure capability can be increased over the time period. It makes no mention of possible increased bomb drop distances through the use of boosted bombs or air-to-surface missiles. The possible mixes of yields and types of the attacker's weapon stockpile can be expected to change over the period. All of these contribute significantly to the threat and must be made a part of our projection of the enemy offensive capability if we are to make a meaningful defense system effectiveness analysis. This figure also does not show the possible attack tactics. These are properly a part of the effectiveness analysis which explores the spectrum of tactics, to see which are the best for the enemy as well as for the defense.

Of course, the same sort of thing must be accomplished if the interest is in an offensive weapon system. That is, to project offensive weapon capabilities we must consider the defensive means at the defender's disposal. The Army's guided weapon philosophy is not different in this case, although the area of operations may be. In the field army's operations the surface support missile system will be attacking a somewhat different kind of target than the Army is defending in the continental United States. If the defensive objective is the air defense of a field army, we must look at the same problem from both sides of the line. We must work the same problem in opposite directions.

Some of the factors from both sides of the problem may be the same. For example, an air defense missile system may have a surface support capability. The presence of an air defense system may limit the tactical mobility of the ground force and in that way influence its nature as a target, and so on.

What is the difference between the effectiveness analysis that predicts effectiveness over time and the one which was mentioned before, the static analysis? This part of the technique is not yet so far advanced as are the two input analyses mentioned. To date, all we have been able to do is to predict effectiveness at successive points in time and draw a smooth curve between the points, which is, I think, adequate. The thing that is needed most is a technique for analysis of the total system effectiveness at each point in time. So far, we are matching capabilities, balancing an offensive weapon system with a defensive weapon system. We are not, in many cases, even considering the complete weapon system, matching only weapon capabilities instead of determining weapon system effectiveness.

This situation is illustrated by considering how effectiveness analyses of air defense weapons systems do or do not include the

weapon control system. The simplest effectiveness study ignores the environment of the weapon. Such an analysis is unsatisfactory for anything but comparison of weapons, but the statement is made that a comparison can be made without significant error. This is not correct. Such a statement can be substantiated only by studying the weapon in its environment and finding out that the effect is, in fact, the same for the two systems. In a surface-to-air missile defense study, the kill potential per unit may be determined under the assumption of perfect, uniform, or random distribution of fire over the attacking aircraft. None of these are achievable in the real case. Also, it is easy to demonstrate that it makes a difference which of these is used, since the comparison between weapons ordinarily is different for each distribution of fire.

If the effectiveness of interceptors is being studied, the assumption of close control or broadcast control is made, with different numbers of expected successes per sortied interceptor resulting from the two assumptions. In this case, as in the guided missile analysis, this is nothing more or less than the substitution of intuition for fact. We have ignored the environment, but we estimate intuitively that the engagement will be this particular way and broadcast control or uniform distribution of fire will give a good approximation to the result. This may be an accurate estimate, but it is a nonverifiable estimate and has a singular disadvantage in that it has substituted the intuition of the analyst for the intuition of the executive, a risky situation at best.

A somewhat better approach to weapon analysis is to assume an environment for the weapon. This approach attempts to state requirements for the weapon control system. We have done analyses in which we say that the effectiveness of a missile defense, using given firing doctrines, has been determined.

The distribution of fire over attacking aircraft is not specified, but the rules for engaging targets are specified. An example of such a set of rules is the "nearest unengaged" doctrine. In this doctrine the first priority target is that target nearest the bomb release point which is not engaged by other batteries; if there are no unengaged targets, then the nearest target is selected. Other doctrines, some more complicated, some less, have also been used.

Now then, what are the requirements which the control system must satisfy if the firing doctrine used is to be feasible? What information, what kinds and how much, must be available to the battery commander in order to carry out the firing doctrine? This includes the information which must be transmitted from battery to battery, from battery to the command or control center, the data rates required, the track capacity which the control center must have, the accuracy with which target position must be measured, and the accuracy with which these data must be transmitted. All these, and more, comprise the list of characteristics which the assumed environment has.

This list can be an input to another study which tries to determine if such an environment could be realized, and if so, when. This study might turn out to show that the assumption was invalid, and herein lies one of the primary difficulties of an analysis of weapon effectiveness which assumes a particular control system. If the study assumes an environment, it is apt to give the impression that the assumed environment could be real environment, and therefore no problem exists. It is somewhat better than ignoring the environment, for the attention of the executive can be directed to the area of uncertainty. He can make his decision with greater confidence in the effectiveness analysis. The analysis may present a set of

results which would obtain for each of several different environments and he can apply his experience and intuition to a clearly stated problem. If none of the control systems indicated can be realized, then the results of the analysis are, of course, meaningless. But the executive and planners at least know that the results are meaningless, a fact which they might not have known if the environment had been ignored.

The kind of effectiveness analysis that is needed, the analysis that is required (if we are really going to be able to leave the decision makers free to make decisions between weapon program alternatives, with confidence that they have a good idea of the consequences of the decision) is one which considers the effectiveness of the whole defense system.

To continue with the example of a weapon and its control system, these must be considered together, as one system. Only in that way can we answer questions such as, "What firing doctrines are feasible using this surface-to-air missile system?" or "How many interceptors can be used with what effect, given the technical capabilities of the SAGE system and the expected traffic densities?" When we have approached defense system effectiveness from this point of view, we will be able to present alternatives which will make much simpler the process of choosing what mixture of weapons is desirable in the defense system and also we can begin to see how to decide which weapons should be phased out and at what times.

The most important advantage which accrues to such an analysis is that the effectiveness analysis can sensibly feed back requirements to the research and development people, requirements which are based on operational effectiveness rather than some idea that this new device would be "nice to

have." It can show when a mismatch occurs between weapon and environment. A missile may be designed with much longer range than it can use, given the limitations of the system to supply target data on distant targets. Similarly, the analysis may show that data accuracy should be sacrificed in favor of higher data rates, or the converse might be true. There are many other such questions, including, no doubt, many that we are not currently aware of. It is clear, however, that the optimum weapon system cannot be designed until the weapon and its environment are studied and designed together. As long as either system attempts to dictate requirements for the other, as is now the case, non-optimum performance is almost certain to result.

This discussion of the influence of the information gathering and weapon control parts of a weapon system applies equally well to the surface support missile case. Such a system has the same sort of dependences as its environment and the weapon characteristics should be influenced by the kinds of information available to the weapon system. A tactical support weapon designed to destroy targets of opportunity cannot be successful in that role unless the information system can locate targets and the system can react fast enough to produce the missile burst prior to disappearance of the target. This kind of consideration may influence the desired maximum range of the system. A weapon of 50-mile range may be nicely matched with the desired mobility of the weapon, the distance behind line of contact that this mobility requires it to be placed, and the quality of intelligence available on targets just behind the battle zone. Of course, the Army always will have use for longer range missiles for use against relatively fixed targets, such as supply dump transportation facilities and the like. But for tactical support, the missile used and its information gathering and processing system should influence each other.

The consideration of the information gathering and processing system, as well as the weapon, is only one of several things which need to be added to the analyses which are currently done to give them the kind of generality commensurate with the scope of the problems which have to be solved. I mention this one in detail since it is one which missile designers have not taken sufficient account of in the past. To produce an orderly weapon system succession, the Army must consider combinations of weapons and systems to meet all the parts of a complex time phased threat.

Let us look now at an example of how such a general effectiveness analysis might come out. Referring to Fig. 6, for the time period in question, the numbers of bomb carriers might look like the upper graph (the same as Fig. 5). The defensive capabilities in terms of surface-to-air missiles might lead to numbers of units available as shown in the middle graph. Both of these two graphs are much oversimplified. The chart does not show countermeasures or altitudes or tactics or many of the other threat characteristics that need to be added to the number of bombers in the attack. Similarly, the defense capabilities shown are only weapons, not weapon systems. Only one of the infinite possible number of different mixtures and phasing schedules is shown.

In any case, as an example, the result of the effectiveness analysis might be the bottom graph which gives the number of bombs on target or fraction of the target system destroyed as a function of the date of the attack. The dotted line might be a tolerable level of damage, that amount of damage we can sustain and have a high probability of being able to pull ourselves together again and fight to win. This kind of plot for a variety of different mixes of weapons and schedules will allow the Army to decide which program it wants to present.

I have not yet said anything about costs. This is not because they are not important, because they may have overriding importance. In fact, because of this, they may operate as boundary conditions. There are a wide variety of boundary conditions which can be in effect in an attack-defense situation of this kind. I will mention only two such constraints and show how costs work in each of these.

One possible framework in which such an analysis might be accomplished is that of stated effectiveness. That is, the systems considered must limit the damage to the target system to x percent, or no more than n bombs are to detonate on target. In such a framework the effectiveness study determines what quantities of the feasible systems are required to satisfy the requirement. Costs of a program to do this might appear as depicted in Fig. 7. This tells the planner how much it will cost him to achieve the desired effectiveness and in what years he will spend the money. Costs of smaller and larger programs will not, in general, be a linear function of these, since effectiveness is not ordinarily a linear function of cost and also because different phasings of programs are required. Another, and much more common to the Army, is the requirement to produce the maximum effectiveness for a limited budget. This is considerably more complicated since costs cannot be determined until after quantities have been programmed over time. The same techniques apply, it is just that the usual way is by successive approximation.

Let us look now at a diagram of how all this analysis is put together in a scheme which assists the Army by providing it continuously updated information needed to put into practice its philosophy of providing continuing effectiveness at minimum cost as shown in Fig. 8. The defense capability analysis must be done in detail for a wide

variety of capabilities; it must reflect the ease of modification of this system and what new capabilities are obtained by this modification. This is the growth potential of the system. The effectiveness analysis will weigh whether that potential is great or small, whether it is valuable or not. The offensive capabilities are treated in the same way and with the same detail, insofar as is possible. Both of these must reflect all possible constraints. An example we have encountered is: "You can have system A this year or system B next year. If you want them both, but as soon as possible, then system A will be next year and system B the year after." This was simply a case of a shortage of research and development capability. A maximum effort on either program would delay the other. Training can be a factor which limits the availability. The training program requires much preparation and it is slow to produce the kind of specialist called for in large numbers in guided missile systems.

The effectiveness matches available systems against the available threat. In doing this, it must talk about complete systems, both offensive and defensive, and if not the entire defense system, at least a complete weapon system. It is no better, however, to match a complete weapon system, environment, mixtures and all, against a partial threat such as manned bombers, than it is to match a complex threat against a partial weapons system. In either case, the result may be quite valuable, but it leaves much to be desired because the unasked questions may have the most important answers.

The effectiveness analysis feeds back to the availability study requirements of two kinds. The first is the knowledge of what time the quantities required will be available. The study will first have been done on the basis of first unit availability and a nominal deployment rate. If the quantities are large,

new conditions may appear. It may be necessary to plan additional facilities; manpower availability may become a limit on the deployment rate. The effectiveness analysis also provides the state of urgency under which the availability is to be predicted. It might indicate that an all-out effort is required to meet a predicted threat. It may indicate paralleling of development operations in order to have high assurance that the project will be ready at the scheduled time.

A second kind of feedback from the effectiveness study to the availability analysis is the requirement to determine when new developments, not previously analyzed, can be made ready to fill a need discovered in the effectiveness study. When can a low altitude solution be deployed? When will the anti-jamming features be incorporated in the radars? There are feedbacks from the effectiveness analysis to the threat analysis also. Analysis may show that some of the offensive capabilities do not pose real threats. The attacker can discover this, too, and will likely redistribute his effort. The modified threat numbers and types go into the effectiveness analysis again. Some attack tactics will prove to be superior. Recognition of this may indicate developments to capitalize on this superiority.

From the effectiveness analysis and the time of deployment indicated by the availability analysis, the annual costs of the defense program may be determined. In the case of a program which is to produce a stated effectiveness, if possible, the annual expenditures may vary widely as new systems purchased are more or less expensive. Note that these costs are not just operating costs, not just procurement costs, but both of these, as well as research and development costs, tooling, testing, training, and all other costs necessary to pay for the defense system. Costs are usually expressed in terms of manpower, dollars, and sometimes amount of

nuclear material. The equations which allow transformation between these are not yet established and must be left to the higher executive and policy making bodies to determine.

The cost analysis feeds back to the effectiveness analysis and the availability analysis whether the specified budget limit has been exceeded and whether the deployment schedules should be stretched out or the total numbers modified. The program may have to be reprogrammed somewhat in order to provide continuing effectiveness over time without exceeding budget limits. A feedback to the threat may appear in the case of a non-budget limited program. If effectiveness is stated, then it may be quite expensive to achieve. Some threats may be easy to meet technically, but the programs to do so may be more expensive than the programs which pose the more difficult technical problems. An example would be the production of very large numbers of bombers which are easy enough to shoot down but which would require very large numbers of missiles to do it. If the attacker had as a goal the long range bankrupting of the economy, he might pick such a course. This kind of consideration is not normally a part of the analysis, but factors bearing on it can be pointed out.

This, then, is how the analysis appears. Only by establishing such an analysis organization and keeping it operating continuously can we hope to be able to make the right decisions. One cannot expect to get all the answers at once. By the time the analysis is done once, it is out of date. However, each successive review is easier and can and should be more general. The first approximation analysis that I mentioned at the beginning could be called the first round through such a loop. Successive trips through this kind of loop are not discrete; the feedback should be continuous and result in a continuous renewal and modification of the

program. New threat parameters may generate new research and development requirements. We must know when these can be met and whether the modification or new weapon can be in time to meet the new threat. We must know what this costs and if we can afford it in the desired amount.

After this process has proceeded for a time, all new weapons and systems become just modifications of the defense system. The arrival of markedly different weapons will have been anticipated long enough that the system can have reacted to the point where the modification is not disruptive. The high obsolescence rate of weapons systems, occasioned by the rapid technical progress in both offense and defense capability, and the long lead times for new equipment caused by the great complexity of modern weapon systems, requires that programs be continually re-evaluated. In so doing, we must be careful to leave open all the alternatives which we might wish to follow up next year. This means

that many more research and development programs must be pursued than before. Most of these will not be followed through to deployed weapons but will be carried along until it is relatively certain that the research and development is not leading to a weapon system that we may need. Alternatives must be left open so that next year's solution is determined by next year's knowledge, not predetermined by the lesser knowledge we have this year.

The object of the weapons program is to provide a continuing effectiveness over a long time. We desire the effect of deterring possible attack and are therefore not pointing for any particular peak effectiveness year. It might be said that the object is to produce weapons systems strong enough that they will not have to be used. With these ideas in mind, it is clear that we must obtain effectiveness by the year, for several years, and continually renew our analysis of these systems in order to get as near the optimum system, defense or offense, as possible.

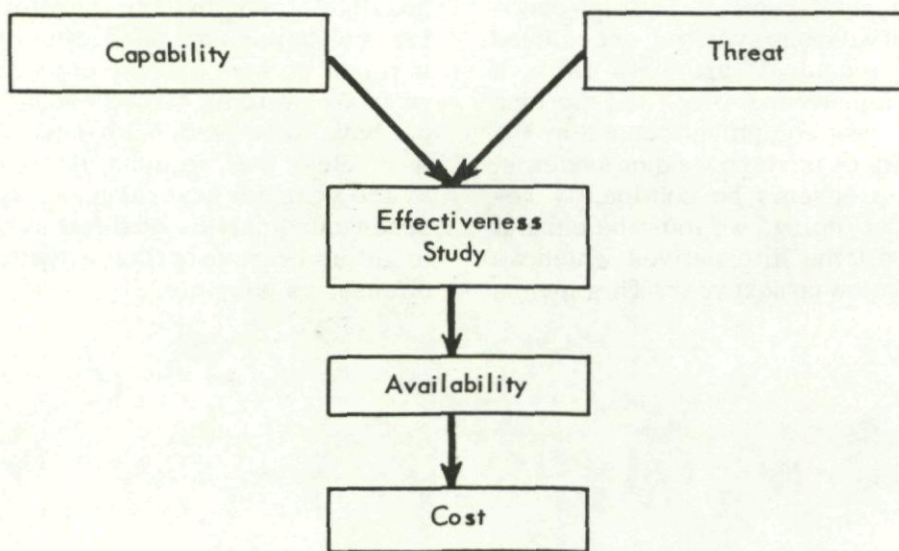
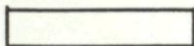
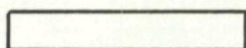


Fig. 1. Simplified analysis scheme.

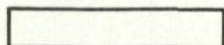
Study



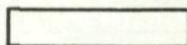
R&D



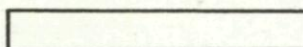
Prototypes and Test



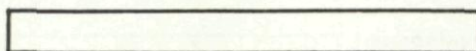
Pre-production



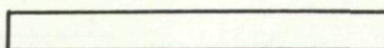
Production



Training and Development
of Tactical Doctrine



Deployment



Operation

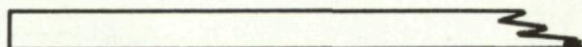


Fig. 2. Lead time phasing.

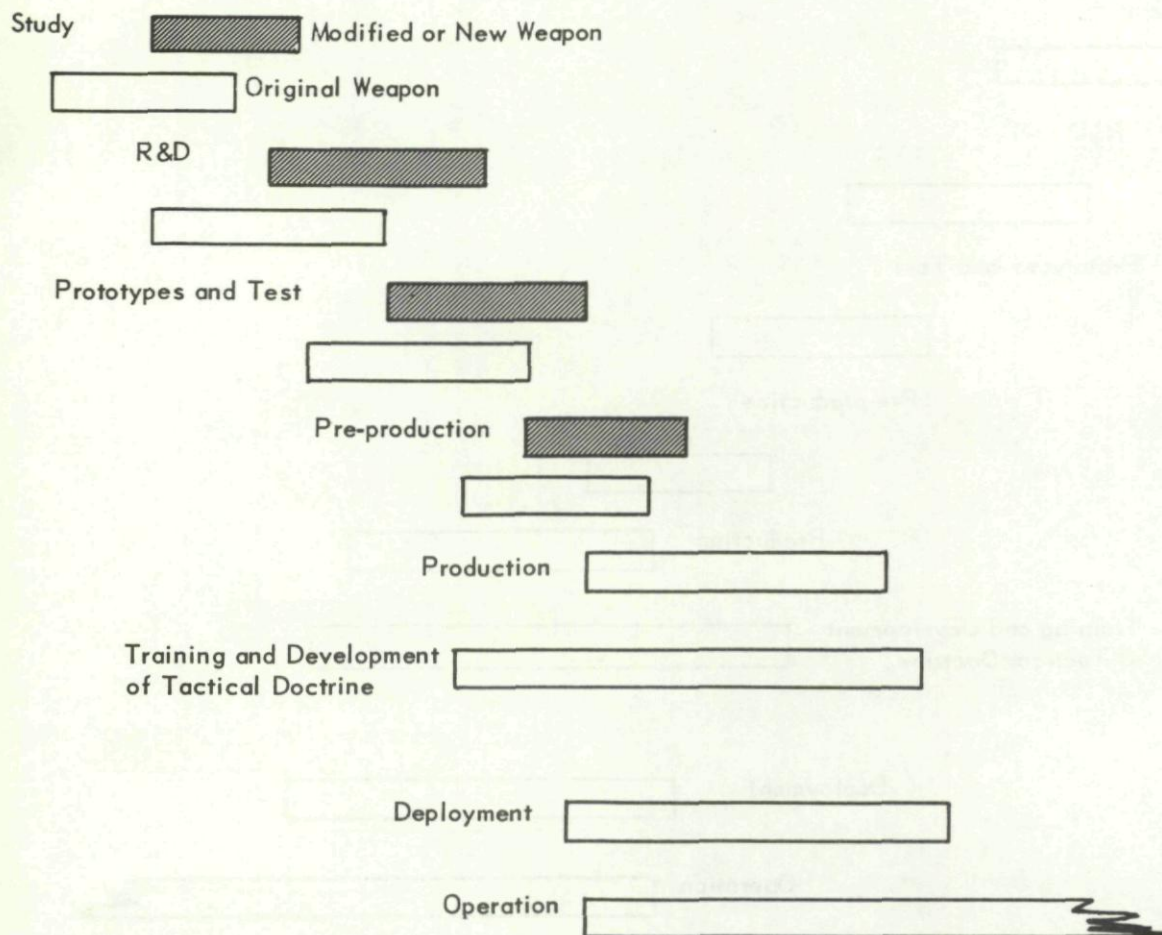


Fig. 3. Lead time phasing.

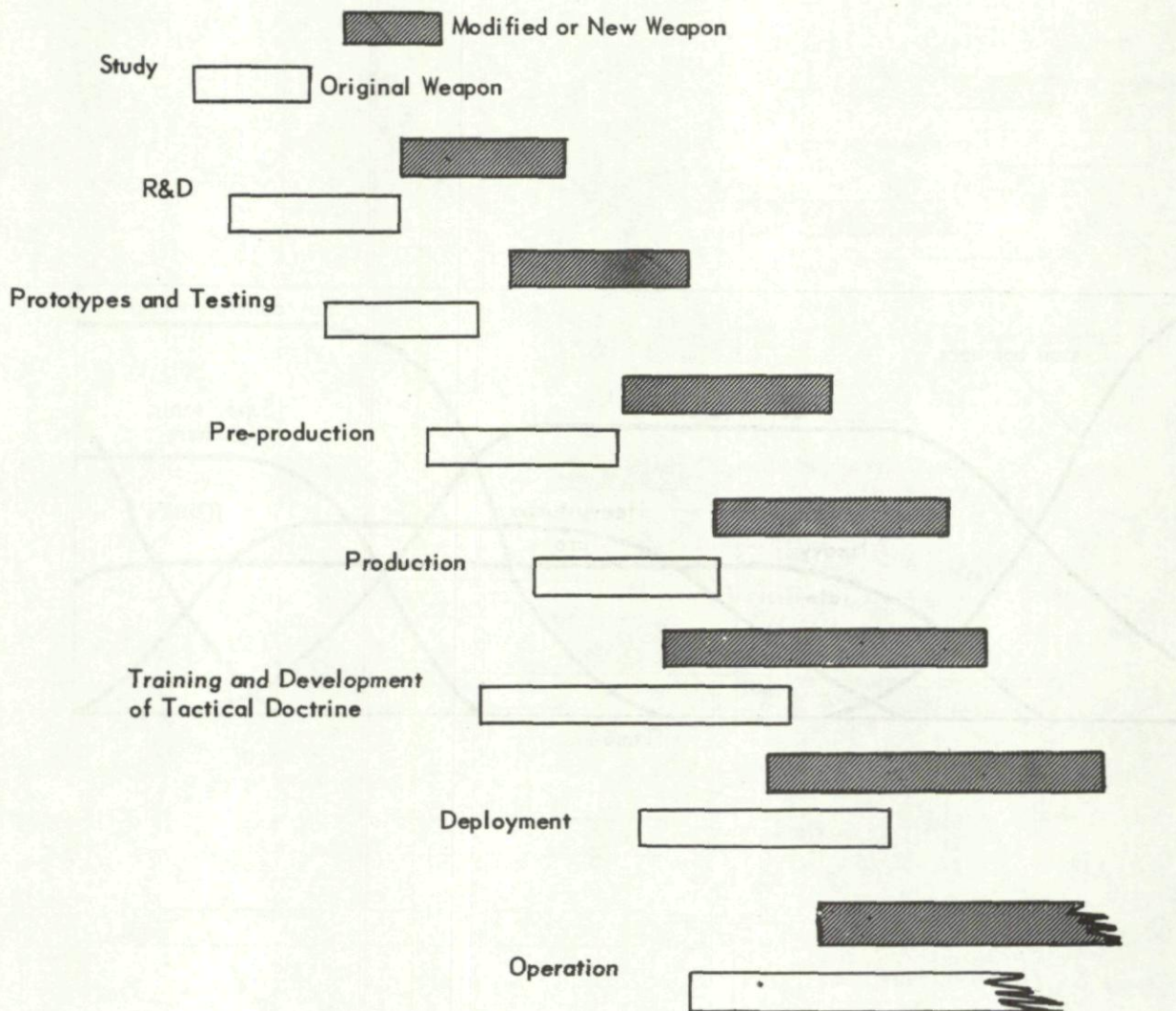


Fig. 4. Lead time phasing.

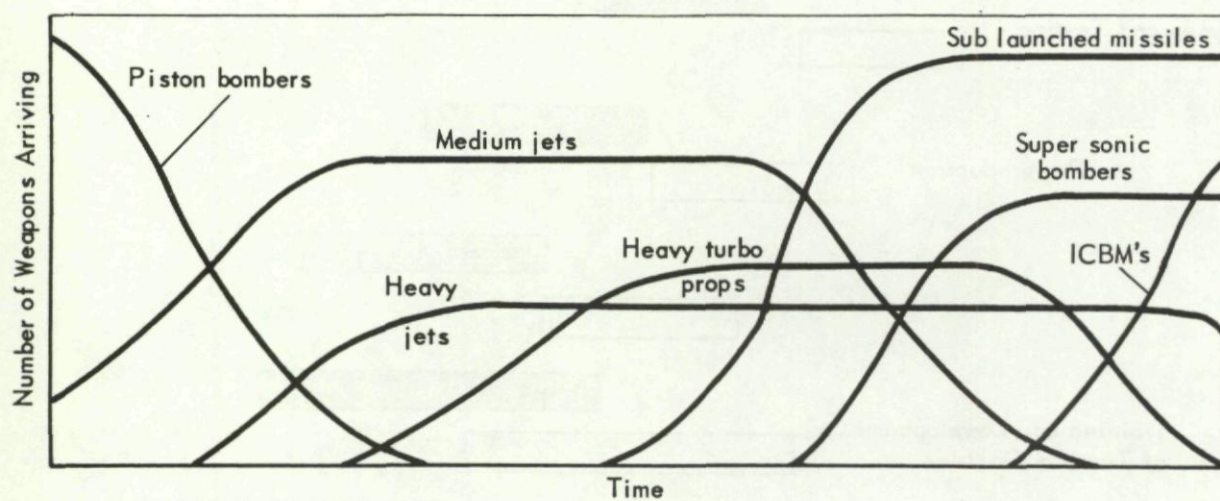


Fig. 5. Availability of complex threat.

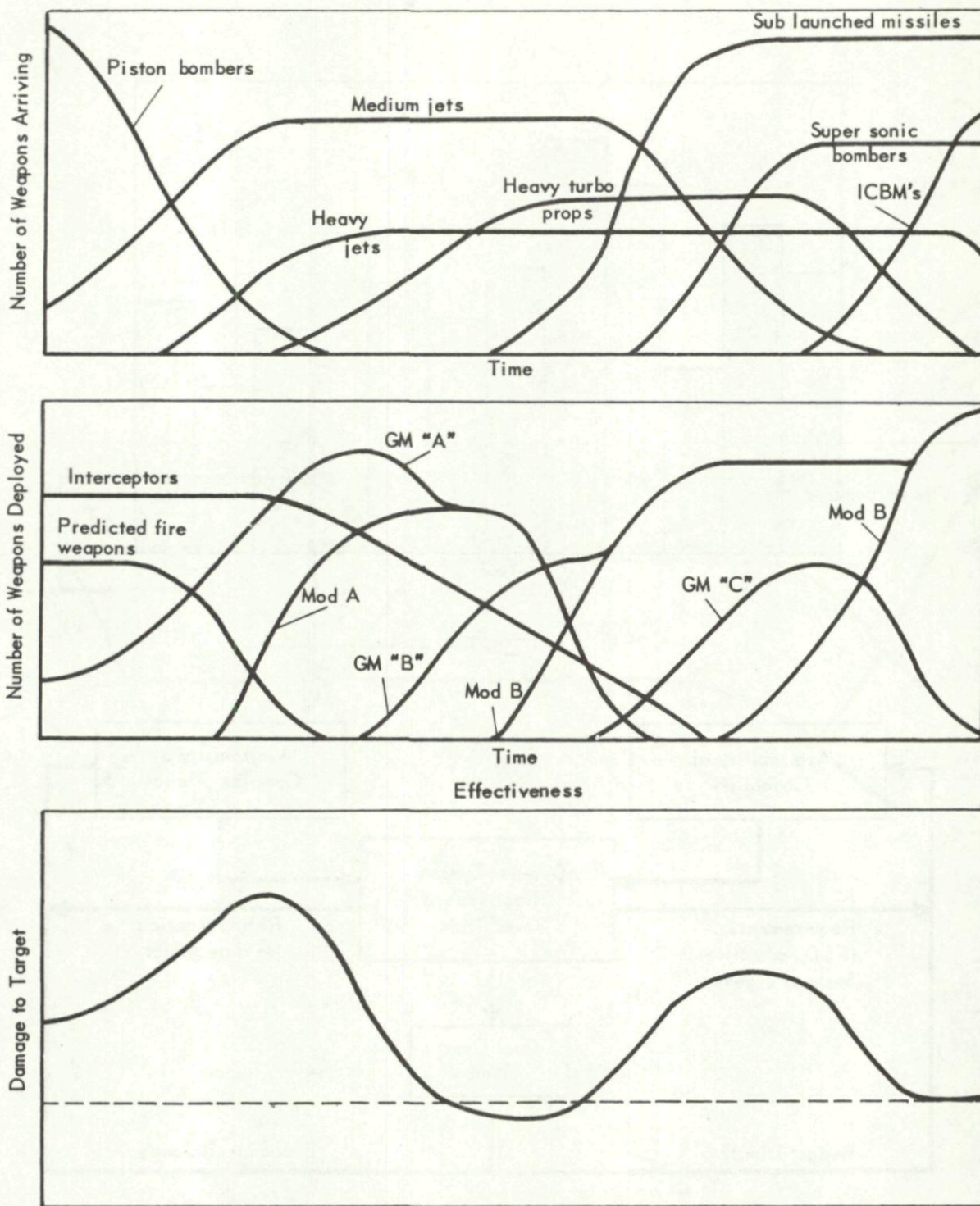


Fig. 6. Top, availability of complex threat.
Middle, availability of defense capability.
Bottom, effectiveness.

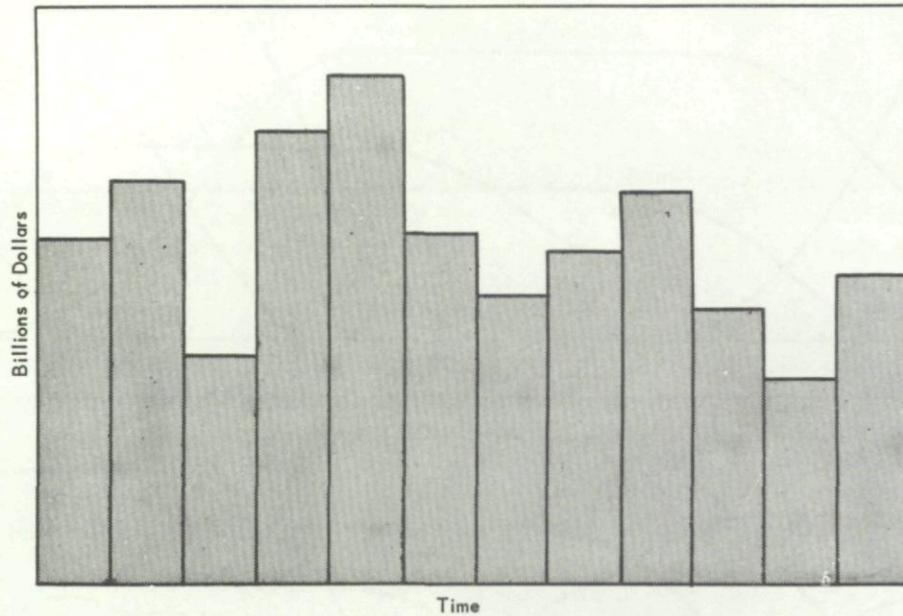


Fig. 7. Program costs.

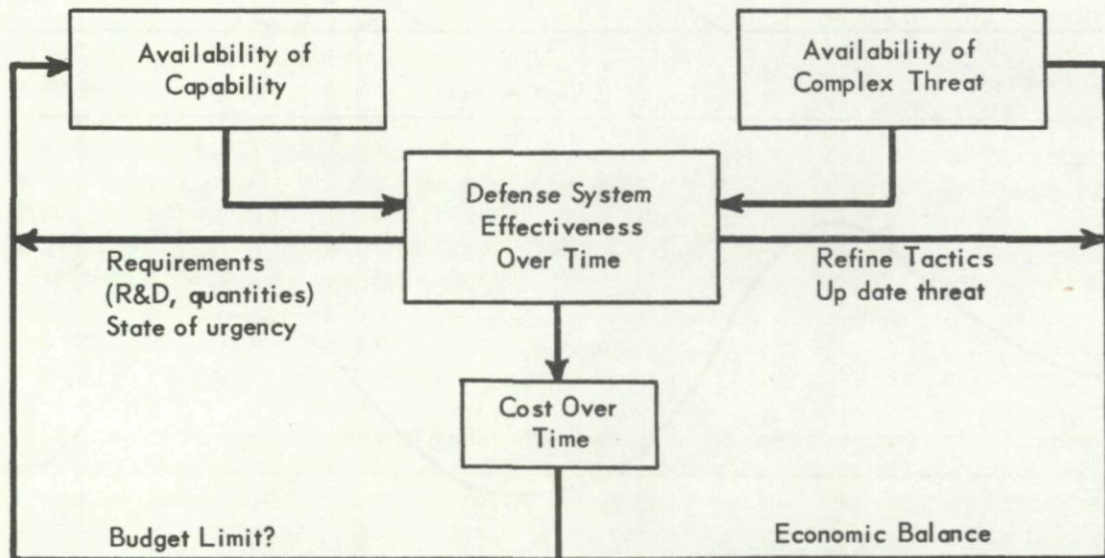


Fig. 8. Analysis scheme.

NEW PRINCIPLES IN THE DESIGN OF SUPERIOR COMMUNICATIONS, NAVIGATION, AND MISSILE GUIDANCE SYSTEMS

W. P. Lear, Sr.*

SUMMARY

This paper deals with the application of statistical concepts to the design of systems. Particular attention is devoted to the operational principles of the SCAN system. The possibility of applying these principles to computers is pointed out.

SOMMAIRE

Cette note traite de l'application de concepts statistiques à l'étude de systèmes. Elle traite en particulier des principes opérationnels du système SCAN et indique la possibilité d'appliquer ces principes aux calculatrices.

1. INTRODUCTION

Communication is generally conceived as the conveying of ideas or symbols from one person to another. In this discussion, we will define communication as the process of transferring messages over substantial distances from mind to mind, mind to machine, machine to mind, or machine to machine. With this definition, navigation and missile guidance systems which accept externally-generated data become special cases of the general communication problem.

Under adverse conditions, the process of communication is a game played against Nature. By exploiting the uncertainty that must exist at the receiving end of a communication system Nature plays a game of augmenting or suppressing the intended message. In radio systems the mechanism for augmenting the intended message is the generation of random signals (or noise) in the atmosphere and in the receiving equipment. Suppression of the message occurs when there are obstacles affecting propagation of a signal between distant points; in

this case forward-scattering, skip, multipath, selective fading, and other anomalies produce uncertainty in the moment-by-moment reception of the radiated signal.

I claim that the vast majority of today's communication, navigation, and missile guidance systems are grossly inefficient and ineffectual compared to what might be achieved. In most cases the equipment limitations are inherent in and arise from the initial statement of objectives. Many systems are inefficient with great power input and extravagant use of the spectrum because the system has not been adapted to its complete environment. We have often played the game against Nature with unnatural rules; for instance, we often design systems which will work only when favorable signal-to-noise conditions exist and we then set about generating the required amount of power, however unreasonable that may be. I would suggest that it is more logical to recognize fully and accept the natural processes and then set about the development of systems which exploit the rules of the game thereby imposed.

*Lear, Inc., Santa Monica, California.

The field of information theory, based upon the classic works of Wiener and Shannon, has formalized a quantitative approach to communication. Of great significance are the theorems relative to the interchangeability of time, bandwidth, and signal-to-noise ratios in such systems. Moreover, information theory has defined such message characteristics as redundancy and coherence, and has indicated their value in increasing the likelihood of message reception in a system which is devised to utilize such properties. The substantial value of integration in signal-to-noise improvement, where the application permits its use, has been demonstrated. Other powerful tools from information theory are autocorrelation and cross-correlation. Experiments have shown the ability of the correlation methods, particularly cross-correlation, to reveal the presence of signals many decibels below the average noise level; here, as with integration, one trades time for signal-to-noise improvement. Certainly these techniques from information theory are important aids in communication system design. However, to this time they have been used mainly in the measurement of the methods we have, rather than to point the way to superior methods.

Additional tools are available from the field of statistics. Measurements made in practical systems always contain some degree of uncertainty. The extent of the uncertainty is in part a function of the measuring unit. For instance, if one made repeated measurements from here to the North Pole with a one-meter rod, a variety of answers separated at one-meter intervals could be expected. In such a simple system the measurements would lie symmetrically about the true value in accordance with the Gaussian distribution. The width of the distribution will vary with the precision of the measuring system and the extraneous disturbances, or noise, added to the process. The significance of any single measurement can only be

expressed in terms of probability. However, in a linear system, the average of a vast number of such measurements would yield an answer of greatly enhanced accuracy. Furthermore, from the field of random processes, it follows that any system error resulting from a very large number of random, uncorrelated error sources will have a probable value which again follows the Gaussian function. These statements suggest means of improving performance in those systems where they can be utilized.

Our present need is to evolve systems which will function under the widest possible range of conditions. Fundamental is the need to devise a method which utilizes all of the available data. To offset the randomness of Nature, it will be necessary to play the game against Nature by placing probabilistic bets, based on sound logic, as to the most likely value of each data sample. For example, in navigation and guidance problems one is dealing with physical bodies having bounded values of original position, velocity and acceleration. Therefore, at any moment all of the possible messages in such a system do not have equal probable value. Why not attempt to find the probable value to each data sample?

2. THE SCAN SYSTEM

My company is currently developing a navigation system which exploits some of the preceding principles and will serve as an illustrative example of the techniques discussed. This system is called the Self-Correcting Automatic Navigator, or SCAN for short.

Navigation systems can be divided into two classes: the self-contained and the externally-referenced. The self-contained class of navigation equipment includes inertial, doppler-supervised, and dead reckoning types. The externally-referenced class

includes all of the radio, radar, optical, and celestial types. The accuracy of self-contained systems is time-dependent. The accuracy of the externally-referenced class is limited by wavelength and line-of-sight or other propagation problems, including signal-to-noise ratios. The error sources for each class of system are different in character.

The SCAN technique is applicable where navigational data from both self-contained and externally-referenced systems are available. Any given self-contained navigational equipment, for instance a dead-reckoning computer, will have measurable error characteristics. A typical dead-reckoning computer is shown in Fig. 1. Aircraft heading and true airspeed measurements are supplied to this equipment. Heading and airspeed corrections due to wind must be inserted manually. Ordinarily such data are not accurately known, thereby reducing the equipment reliability. The SCAN system provides the means for automatically and accurately learning these wind values. The corrected velocity vector may now be converted from its polar form into north-south and east-west components. By integrating each velocity component with respect to time, one obtains the distance traversed in that ordinate. By adding the original coordinate position, which is the constant of integration, one obtains present position.

A typical error characteristic for such an equipment is shown by the curve in Fig. 2. The curve depicts an error versus time contour which will contain almost all statistically observed errors; say 99.5% of all observations fall within this contour, which we shall call the confidence limits of the self-contained system. Therefore, after operating over an interval t_x without recalibration, errors as great as plus or minus ϵ_x are acknowledged as possible. However, since the total error results from the summation of

several uncorrelated sources (in the dead-reckoner these are the random errors introduced by the directional reference, the airspeed transducer, and the coordinate converters and integrators) the probability of various error magnitudes follows the Gaussian function. This indicates greater likelihood of the occurrence of small errors than of large ones. The cross section of the confidence limit can therefore be said to have a Gaussian profile. The existence of a confidence limit, based upon physical measurements made on a particular set of hardware, is normally not recognized as meaningful data and is not utilized in system design.

If one examines the manner in which externally-referenced navigation systems are used at present, for example a radio system like VOR-IME or Tacan, we find a most peculiar form of data processing. First of all, the observer must either completely believe or completely disbelieve what the equipment indicates. If no other navigation data are available, belief in the one set of data is a pure act of faith. The observer, however, is encouraged in such faith when the dials or indicators are relatively stable and moving at a reasonable rate. This stability is a measure of signal-to-noise. Therefore, when favorable signal-to-noise ratios exist, confidence is high; but when poor signal-to-noise ratios are encountered and dials show excessive wander, all confidence is lost, even though the equipment may on frequent occasions be indicating valid data.

It may now be apparent that the confidence limits and the associated Gaussian probability function which applies to the self-contained system can be used in evaluating each sample of data from the externally-referenced system. By weighting each data sample with the Gaussian function all sample points outside the confidence limits are given

a weight of zero and are automatically discarded. All points inside the confidence limits, whether noise or valid data, are given a Gaussian weight and are then integrated. Since noise has no preferential position, over an interval it will integrate toward zero value. However, if a valid error exists and even if only occasional data samples are received, such information is not a random process and will result in a definite integrator output. The existence of such positional errors, that is differences between the self-contained and the externally-referenced indications, is used in a manner which causes the two systems to come into better and better long-term agreement.

Fig. 3 shows the manner of converting the example dead-reckoner into a system which accepts and weighs outside data. Observe that from the initial inputs, heading, and airspeed, down to present-position data this is still the same unit previously shown. Provision is made for accepting data from the radio system and subtracting these from the local data. Obviously, identical coordinates and units must be used. The "data filter" shown in the figure is a functional unit which embodies the confidence limits and applies the Gaussian weighting. The output from this unit, which is probability-weighted error, is then integrated. This integral value can be shown to be the velocity correction required in order to cause the locally measured velocity to come into long-term agreement with the external data. Furthermore, since the external data are measured against a ground reference this velocity correction converts the local measurement from airspeed and heading to ground speed and ground-track. This correction is therefore the wind effect (if all local measurements were error-free) and is injected at the dead-reckoner wind inputs.

Fig. 4 illustrates the underlying philosophy of the SCAN system. Here data from a self-contained and an externally-referenced

system are compared or subtracted one from the other. The difference will contain the noise of both systems plus any valid error data. If one had an optimum filter (in this case a filter capable of accepting valid data and rejecting noise) the output of the filter would consist of only such corrective data as required to make the two systems agree. The practical filter, which approaches the optimum in practice, consists of the statistical processing of the data filter and the functioning of the double integrating closed-loops shown in Fig. 3.

In practice, the measured statistical drift characteristics of the self-contained system are used to establish a confidence-limit circle about the indicated position of the vehicle. The radius of the confidence-limit circle depends upon the interval of time since acceptable corrective data were received and the system drift rates. Radio position data samples are individually given a probability or credibility coefficient determined by the relationship of the indicated position and the confidence limit. This constitutes a multiplication of the radio system distribution by the self-contained system error probability. This product, which is the probability-weighted error, is integrated to provide true cross-correlation. Long-term integration further constitutes the taking of a large statistical sample; no conclusions are ever drawn on one, or a few, samples of data.

The special purpose machine thus devised "knows" how to maximize the correlation and in so doing learns the winds aloft and causes the two sets of navigational data to come into closer and closer agreement over the long-term. Furthermore, it can obtain added correction data even with sporadic radio signals or adverse signal-to-noise ratios. It can provide velocity memory for accurate extrapolation during intervals when external data are unavailable and knows how to pass judgment on data when signals are restored.

Because of the system's sophisticated data filtering, it is relatively immune from interference effects from either random or specifically generated signals.

The SCAN system just described may serve as an example of my broader thesis; there are, however, no set rules for the development of optimal systems. Every aspect of the problem must be explored; every bit of known or implied information must be utilized. Often the setting of the problem must be adjusted in order to permit the use of best techniques. For instance, in the SCAN system long-term integration became possible only when the terrestrial coordinates were changed into a moving set of coordinates attached to the vehicle.

Most information sources are high in redundancy of data; few systems make effective use of this added information. In general, any system which simply displays currently received signals, which does not use storage or does not at least know the recent history of the data, can be said to be nonoptimum. This applies to all applications except those which transmit completely random messages or numbers.

3. COMPUTERS

As one further, and more speculative, example of this approach to system design, I would like to discuss some unorthodox thoughts about computers. The computer may be regarded as a communication problem wherein only the element of distance has been eliminated. The ordinary approach to general-purpose digital computer design is characterized by the sizable tonnage and kilowatt ratings of the end product. Such equipment results from a design objective which seeks equipment infallibility. This objective enforces prodigious signal-to-noise ratios and multiplication of complexities due to the

additional self-checking and other safeguards. Each of these steps further increases input power and size. Infallibility itself is often an elusive goal, since the steps taken to achieve it often, of themselves, reduce reliability.

Now, following the principles previously discussed, let us consider the application of statistical processes to such a computer. First of all, let us say we will use transistors and extremely low power input. In fact, let us work with signals at or near the noise level, so that the resultant probability of getting a correct answer to a problem may be only ten percent. What good is a computer with ten percent reliability? Well, let us acknowledge that if we knew how to design such a computer it might be made vastly smaller than its big, reliable counterpart. Perhaps it could be made to occupy only one-third of a desk-top and be powered by a single dry battery.

Now let us attack its value as a computer. What if we chose to use three such computers (still very, very much smaller than its counterpart) and connect them so that problems were applied to all three in parallel. Let the answers out of each computer then go to a small programmer which makes a coincidence check on the several numerical results. Only when all three machines agree will the answer be accepted; until they all agree on a common answer the programmer continually reinitiates the same problem. Eventually such an agreement will occur and the answer will come out. If simultaneously correct answers are required, the probable number of computations (for ten-percent reliability) is one thousand. However, if some memory capability is included in the design and nonsimultaneous agreements accepted, the probable number of computations can be considerably reduced.

If the computer in question had as little as ten thousand output numbers, each of which were equally likely to be noise induced, the probability of a mutually agreed upon wrong answer is less than about 10^{-9} *. This is a very respectable reliability, even for the so-called infallible machine.

You may now be prepared to ask why three computers, why not ten or two or one? Ten or two will only change the statistics of reliability. One machine with three

*This comes from $(10^4)^3$ or 10^{12} for one calculation; however, since on the average at most 10^3 calculations are required, this figure must be divided by 10^3 .

computations may seem the equivalent, but is not. The statistical process will provide signal-to-noise improvement only if the error processes are random. This cannot be assured in a single system; consider specifically a component failure which would consistently tend to repeat a given error.

If there are any computer designers reading this paper, I offer my sincere apologies for any apparently disparaging comments I may have made. In fact, I stand in the utmost awe of their monumental but capable machinery. It is my hope that this computer speculation does serve to dramatize the possibilities and methods of statistical design. Too long, I think, have we tried to live in a black and white world, populated only with ones or zeros for probability.

REFERENCES

1. Shannon, C. E., and Weaver, W., "The Mathematical Theory of Communication," University of Illinois Press, 1949.

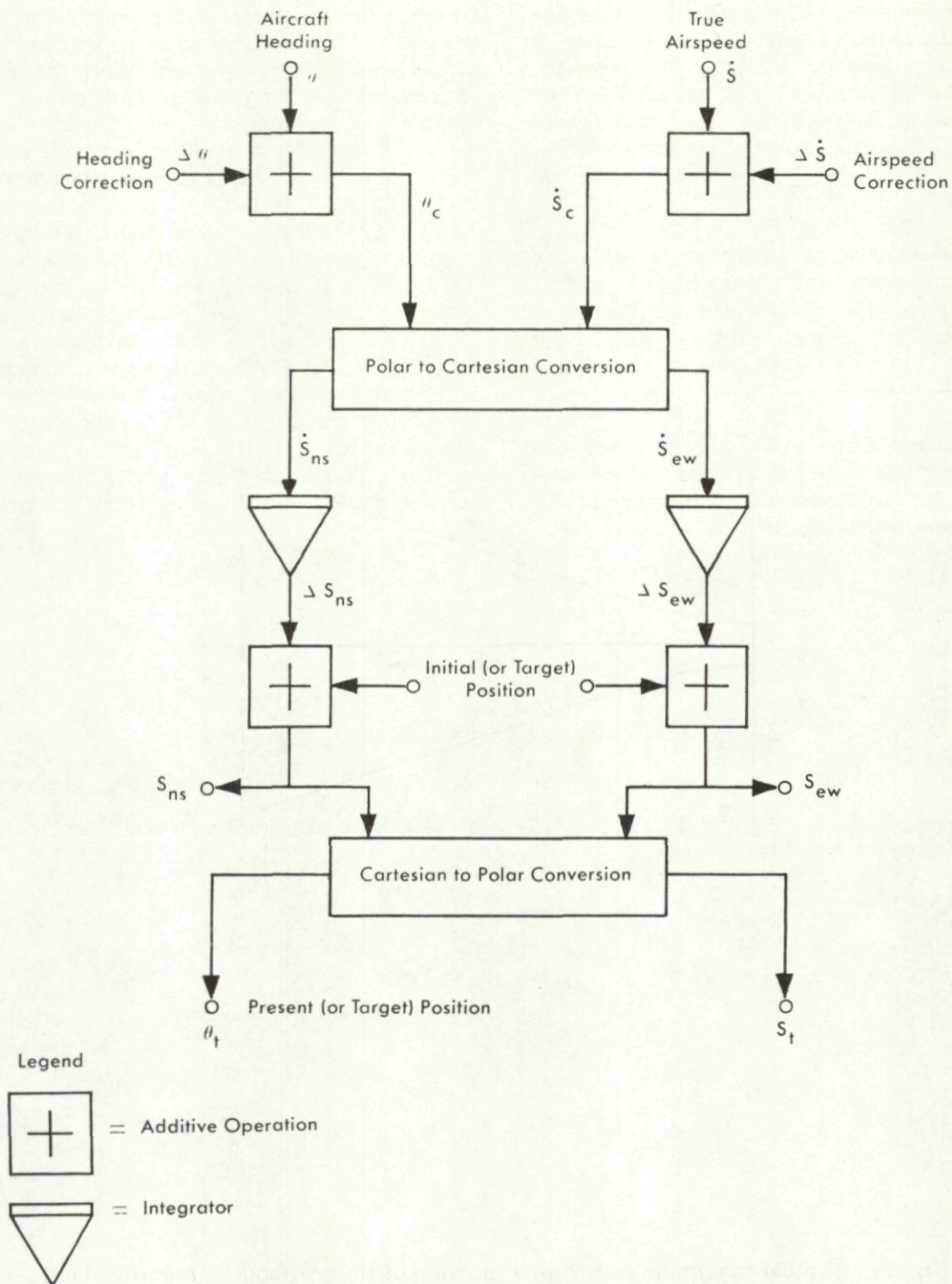


Fig. 1. Block diagram of simple dead reckoner (open loop).

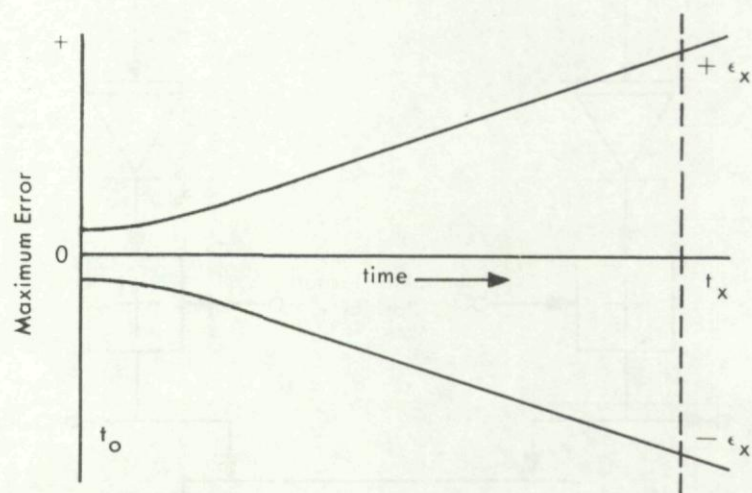


Fig. 2. Typical maximum error as a function of time without correction data.

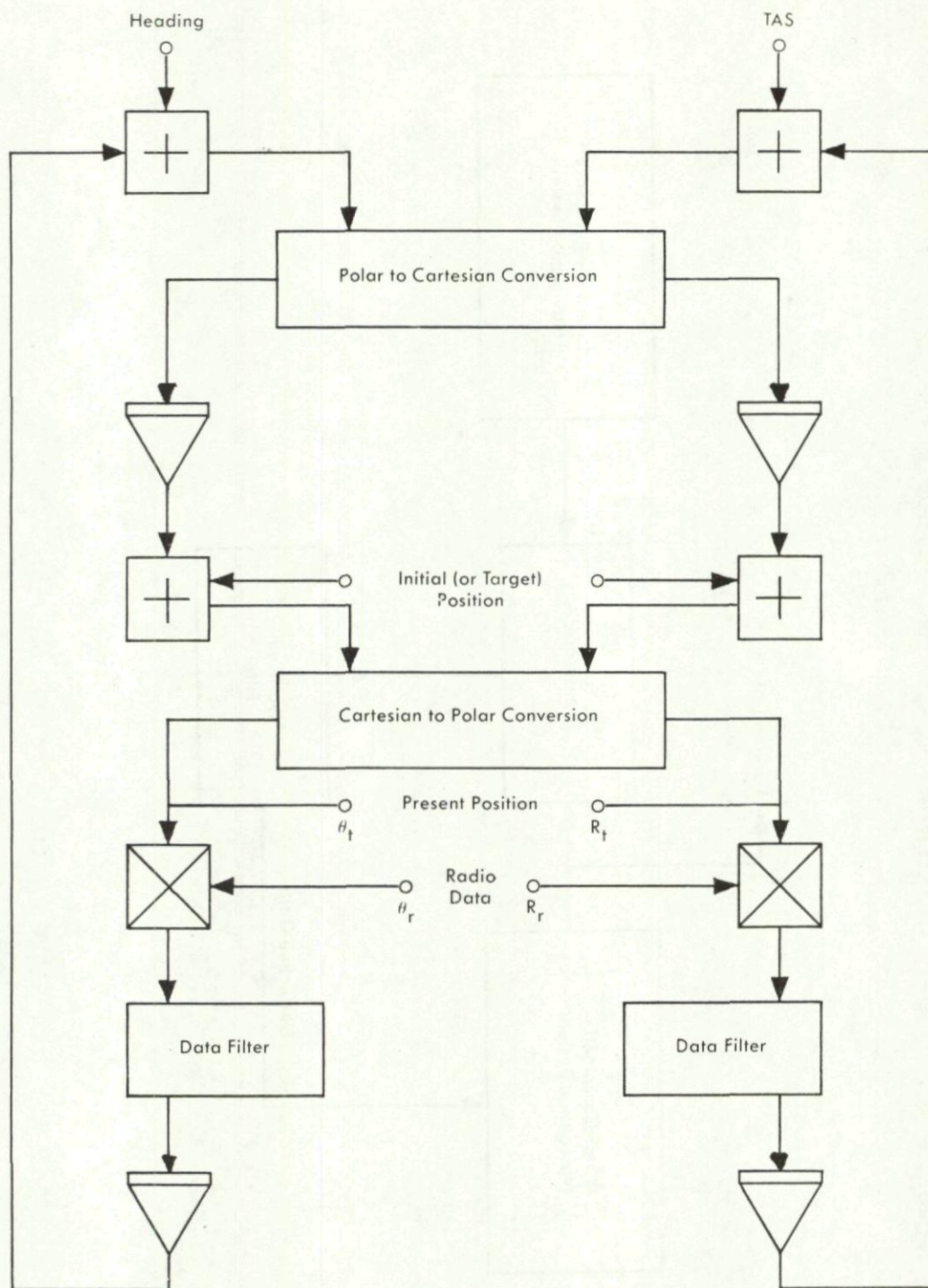


Fig. 3. Block diagram, self-correcting automatic navigator using heading and true airspeed with polar radio data.

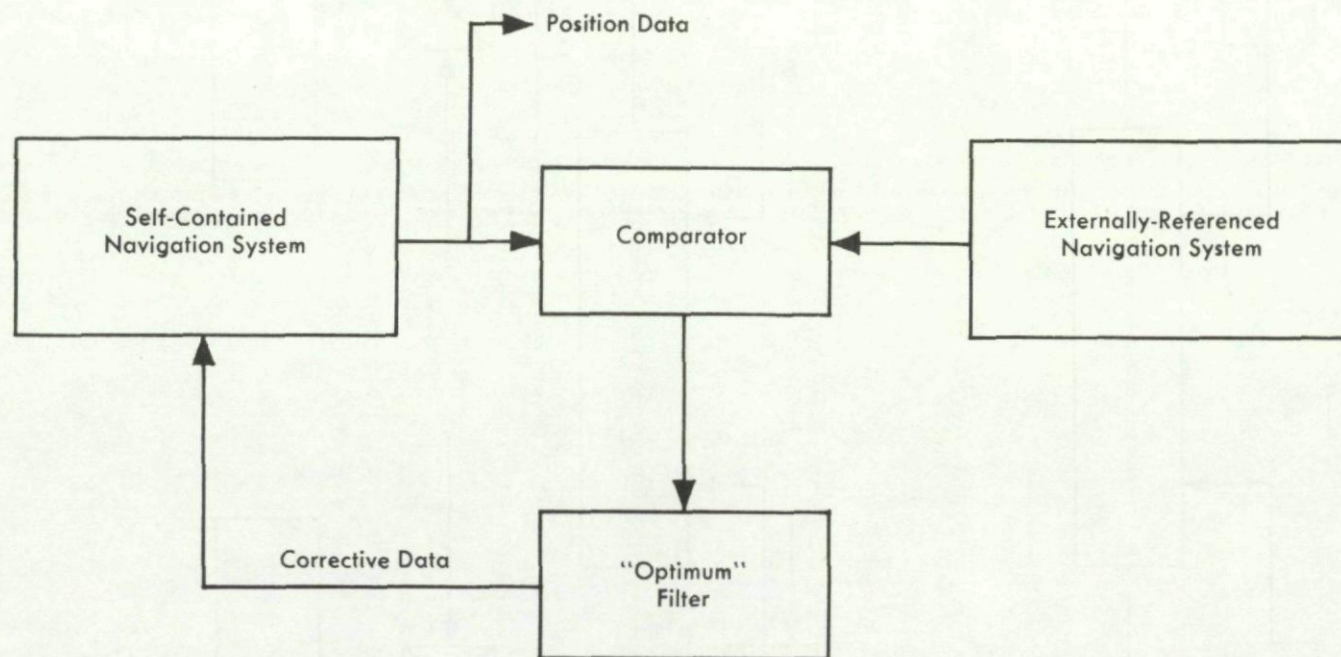


Fig. 4. Simplified diagram of SCAN method.

GUIDANCE TECHNIQUES

Walter L. Webster, Jr.

SUMMARY

There are a number of possible guidance techniques which can be used to guide a missile toward a collision with its target. These are command, beam rider, radio navigational, homing, and combined techniques, each of which has its advantages and disadvantages. The command system requires complex ground equipments and guides the missile by continuously monitoring the position of both the missile and the target. The beam rider follows a beam which irradiates and locks on the target from the missile source. Radio navigational schemes are very accurate in positional placement of the missile but require a ground baseline installation. Homing techniques rely on reflections or radiations from the target and have the advantage of having higher and higher signal levels available as the missile approaches its target. The choice of system depends on a number of factors such as the accuracy, the medium from which the missile is launched, the number of simultaneous targets and missiles which must be handled, and the range over which the missiles must be effective.

SOMMAIRE

Il existe un certain nombre de techniques de gouverne possibles qui peuvent être utilisées dans la gouverne d'un missile vers la collision avec sa cible. Ceux sont: les techniques de commandement, de suiveur de rayon, de radio navigationnelle, de "rentrée" et les techniques combinées, dont chacune a ses avantages et ses desavantages. Le système de commandement exige des équipements au sol complexes et il guide le missile en relatant continûment les positions respectives du missile et de la cible. Le suiveur de rayon suit un rayon qui irradie et se fixe sur la cible à partir de la source du missile. Les dispositifs de navigation par radio sont très précises dans le positionnement du missile, mais ils exigent une installation bien déterminée au sol. Les techniques de "rentrée" s'appuient sur les réflexions et les radiations émanant de la cible et elles ont l'avantage d'obtenir des signaux d'entrée dont le niveau est de plus en plus élevé à mesure que le missile s'approche de sa cible. Le choix d'un système dépend d'un certain nombre de facteurs tels que, la précision, le véhicule duquel le missile est lancé, le nombre de cibles et de missiles qui doivent être simultanément maniés et les distances pour lesquelles les missiles doivent être efficaces:

1. INTRODUCTION

A missile guidance system has been defined as a group of components which measures the position of a guided missile with respect to its target and causes changes in the flight path as required. A subject as

broad as this, one which encompasses so many of the engineering and basic sciences, cannot be treated in detail in the length of time allocated to any one speaker. I will try, therefore, to cover with a rather broad brush the various types of guidance systems with which I am familiar, their advantages and

disadvantages, and some of the associated engineering problems which we in the United States have experienced in the design and testing of such systems during the past decade.

In the usual case, the guidance system will include sensing, computing, directing, stabilizing, and servo components. Not all of these components need be incorporated in the missile and in fact it will be seen that in a majority of the practical cases a portion of these components is external to the missile and located at the launching site. The stabilizing and servo components are always located in the missile and are often referred to as the control system, although they are considered to be component parts of the overall guidance system.

The objective of a guidance system is to deliver a missile at the target with sufficient accuracy to accomplish its mission. In the military sense the mission is usually target destruction. Missile targets are generally classified into two broad groups, air targets and surface targets. Surface targets may be further broken down into stationary targets and moving targets; I think you will agree that all air targets are moving targets. It will be found that the above categories of targets will have a great influence on the type of guidance system which is chosen for any given missile system.

At this time I would like to outline five broad categories of guidance into which most known systems will fall:

- (a) Command Guidance
- (b) Beam-Rider Guidance
- (c) Radio Navigational Guidance
- (d) Homing Guidance
- (e) Combined Guidance Systems

2. COMMAND GUIDANCE SYSTEM

A command guidance system is defined as one in which guidance intelligence transmitted to the missile from a remote source causes the missile to follow a directed flight path. Inherently, command systems are the simplest of all the techniques considered. Command guidance systems have long been used for the remote control of boats, tanks, model aircraft, and target aircraft. In general, command systems require that the behavior of the missile and the target be monitored, that a prescribed collision course be computed, and that information be transmitted to the missile control system so as to align the missile flight path toward an intercept with the target. Its most serious general limitation for military missile use is low traffic-handling capacity, for the typical command system will solve only one intercept problem at a time. Owing to the fact that the behavior of the missile and the target must be monitored and that this is most often accomplished by tracking devices located at the launch site, another serious limitation is its decreasing accuracy with increase in range.

The principal functions of a command guidance system are the acquisition and tracking means, the computation of flight path error, communication of the intelligence to the missile, and a sensing and control mechanism in the missile to accomplish flight path correction.

A typical command system would look functionally somewhat as shown in Fig. 1. This system provides for separate tracking of the missile and target in bearing and range, and a separate command link. There are, of course, many variations of this type of system and it may be used for air-to-air, air-to-ground, and even surface-to-surface missiles as well as for the surface-to-air case shown here. The obvious advantage of such a system is derived from the fact that

most of the complex equipment is located at the launch site instead of in the missile. The missile carries only a bare minimum of electronics - a command receiver and stabilization and control equipment, thereby increasing its reliability and decreasing missile cost.

The design of the launching site equipment for this system is conventional, its complexity depending upon the specific application. In the case shown, the tracking requirements for missile and target are practically identical, while in the case of the surface-to-surface missile where the target is fixed there will be no requirement for target tracking, and the target coordinates can be set into the computer. The computer design again will be straightforward but will vary in complexity with type of flight path solution required and the number of additional functions it is called on to perform, such as parallax computation, pointing of the launching platform, etc.

Since communication of the guidance intelligence to the missile is most often accomplished by radio link, it will be found that in the nonmilitary case this component is the least difficult to design of all the elements of the guidance system. In general, UHF and VHF frequencies will be used to take advantage of the reduced size of components required in the missile, although the controlling factor will be the range at which the missile system will be required to operate. In the military application, command link security will be a design consideration of utmost importance. Ideally, the link must be immune to interference from identical systems operating in the same or adjacent areas, interference from communication and navigational systems of friendly forces, and from detection, analysis, and jamming by the enemy. Such considerations greatly complicate the problems involved in the design of a command link for military application.

3. BEAM-RIDER GUIDANCE SYSTEM

A beam-rider guidance system is defined as a system for guiding missiles which utilizes a beam directed into space, such that the center of the beam axis forms a line along which it is desired to direct a missile. (The most practical type of such a system is that using radar to form the beam.) The missile contains equipment that determines the direction and magnitude of the error when the missile deviates from the center of the beam.

Fig. 2 shows a simple diagram of a beam-riding missile guidance system when a radar beam is used as the guiding medium.

A radar antenna is directed so that the center of the beam is on the target. Information is superimposed on this radar beam which permits the missile to measure its position with respect to the center of the radar beam and to move in a direction which will reduce this error toward zero. The example shown is an air-to-air application; similar systems can be devised for surface-to-air and air-to-surface application. It will be seen that in all applications of the beam-rider principle, there must be a line-of-sight path between the guidance radar and the target at all times.

One requirement of the system is a high order of accuracy in placing the radar beam on the target. In most cases the radar will be an automatically tracking radar in which case a precision tracking radar with high angular accuracy, such as is used for airborne intercept or shipboard fire control, will meet the requirements.

Mention has been made of the necessity for the beam to contain information which will permit the missile to orient itself with respect to the center of the beam. The guidance

information must contain at least two facts: the amplitude of the error and the direction of the error. With this intelligence the missile can correct its error in position measured from the center of the beam.

In the conventional scanning radar the center of the radar beam is pointed at an angle from the scan axis and the beam is rotated so that the center of the beam follows the scan circle as shown in Fig. 3.

If a receiver is placed anywhere on the scan axis, the signal strength measured by the receiver will be constant as the beam circles around the axis. If the receiver is displaced from the axis the amplitude will be modulated at the scan frequency. Fig. 4 shows the varying amplitude modulation which results from placing the receiver at three different points with respect to the scan axis.

When the receiver antenna is displaced from the scan axis at (a), a sine wave modulation appears in the amplitude of the received signal with the phase of the modulation, such that the maximum amplitude appears at 90 degrees. When the receiver is at point (b), the modulation is still present but reduced in amplitude and the phase has changed so that the maximum appears at 180 degrees. At point (c), the received signal is constant.

It will be found that as long as the receiver is maintained near the scan axis of the radar, the signal amplitude modulation will be directly proportional to the displacement of the receiver from the center of the beam. A beam radar receiver, therefore, already has a measure of the amount of error of the missile's displacement. It has already been pointed out that the amount of error is not enough, but that an additional fact must be supplied to the missile; namely, the direction of the error. The phase of the frequency modulation is available in the missile receiver and needs only to be compared with a reference modu-

lation at the radar scan frequency to disclose the direction of the error. The reference modulation may be obtained from the reference generator already existing in the radar and may be transmitted to the missile receiver in any manner the designer may select. If a suitably stable oscillator can be designed, the reference generator might even be carried in the missile and synchronized with the radar just prior to launch. After the signals are received in the missile, they are separated into a voltage proportional to the displacement of the missile from the scan axis of the radar and a voltage proportional to the angular position of the missile in the radar beam as measured from some arbitrary zero reference. The combination of the two signals represents in polar coordinates the error of the missile from the center of the beam. This information may be converted into rectangular coordinates to control the missile in pitch and yaw.

In a beam-rider system the equipment carried by the missile is more complex than in the simple command system previously described, but less complex than most of the systems to follow. The overall system is less complex than the typical command system, since only one tracking radar is required and computer requirements are much less severe. Although a beam-rider guidance radar is required for each target being attacked, more than one missile may be guided simultaneously by the beam. The problem of mutual interference between similar guidance systems being used in the same area exists in this type of guidance but is far less severe than in the previous case.

4. RADIO NAVIGATIONAL GUIDANCE

Both circular and hyperbolic navigation techniques are applicable to missile control and various forms of each have been used for years in the control of piloted aircraft. They are primarily applicable to surface-to-surface and surface-to-air missile systems.

Hyperbolic navigation is a general method for determining lines of position by measuring the difference in distance of the aircraft or missile from two or more stations whose positions are known. By measuring the difference of time of arrival of signals transmitted from the stations, the distances can be determined.

A typical hyperbolic guidance system is shown in Fig. 5. A master transmitter is carried in the missile and periodically transmits a signal which is received by each of the ground stations. Each station then transmits a signal back which is received by the missile. A time comparison of the two signals is made in a comparator and from this the necessary time or distance difference is derived. The missile can then be made to fly on a hyperbolic path which represents a constant distance difference and which passes through the target. Altitude is controlled by an altimeter. Distance from either of the ground stations can be measured independently by measuring actual round trip time of transmission of the signal and when this distance equals the distance from the station to the target, the terminal dive can be initiated.

Hyperbolic systems can be designed to give excellent accuracy. Their traffic handling capacity is good and many widely-spaced launching sites could be used to launch missiles into a single hyperbolic guidance pattern. The missile does not have to be tracked, as in the command type system, although the tactical situation may make this desirable for information purposes.

A very similar radio navigational type system, differing primarily in geometry is the so-called circular navigational system shown in Fig. 6. In this system the missile flies along a path of constant distance from a ground transmitter describing, of course, the arc of a circle. If this constant distance is

equal to the range of the ground station from the target, then the arc along which the missile flies will pass over the target.

The advantages of such a system are identical with those of a hyperbolic type system and the same tactical considerations apply; namely, the target must be a stationary one and the distance of the target from the two slave stations must be known. It should be noted that in both these systems, only distances are required and that the distance between the two slave stations need not be known exactly.

5. HOMING GUIDANCE SYSTEMS

A homing guidance system may be defined as a guidance system by which a missile steers itself toward a target by means of a self-contained mechanism which is activated by some distinguishing characteristic of the target. There are three general types of homing systems; i.e., active homing, semi-active homing, and passive homing.

A passive homing guidance system may be defined as a homing system wherein the missile makes use of energy emanating from the target. In such a system no transmitter or illuminating source of any kind is required in the missile, but merely a receiver which is capable of detecting and tracking the particular type of energy being radiated. This energy may be in the form of light, heat, sound, or electromagnetic radiation.

Of the three basic homing systems, the passive system requires the least equipment in the missile and, therefore, holds out the promise of greatest reliability. The general guidance technique for visual light or infrared homing guidance is the same, differing only in details of the sensitive element and the system of optics. A sensitive element, an optical system to focus the energy on the element,

and a method of scanning such as is used in radar systems is required. By commutation, the amplitude of the infrared signals from four quadrants may be compared and up-down, left-right signals generated for missile control. An obvious drawback of the infrared or optical guidance system is its poor performance under adverse weather conditions; obviously the operating range of such a system will be reduced considerably by haze and rain and will approach zero under conditions of heavy fog and rain. Furthermore, since the sensitivity of the systems depends upon the contrast of the target with its background, those targets which do not emit light or heat sufficiently different in intensity from their background will make the system useless.

Radio frequency passive homing systems are best exemplified by their relatively simple counterpart in ships and aircraft navigation, the radio direction finder. The requirements are for a receiver which can be tuned automatically to the target transmitting frequency and an antenna and signal comparison method to give the required directional information. Both phase comparison and amplitude comparison techniques can be used to derive this directional information. The major problem involved in the design of such a system is the requirement for tremendous dynamic range in the receiver, since it must operate on a signal which is relatively weak at long range and very strong as the missile approaches close to the target.

An active homing system is one in which both the transmitter of energy required to illuminate the target and the receiver of the reflected energy are located within the missile itself as shown in Fig. 7.

The energy used to illuminate the target may be in the form of light, heat, radio, or sound, the most common form being the

radar seeker. While theoretically either CW radars using velocity measuring techniques or pulse radars using range measuring techniques can be used equally well in this application, it will be found that certain fundamental characteristics of the CW radar make it very difficult to engineer into an active missile seeker. Since the receiver and the transmitter of the CW radar must be on at the same time, the receiving and transmitting antennas must be physically or electrically separated. In the practical case the antennas are physically separated by a great enough distance to eliminate the problem of feed-through of transmitter energy directly into the receiver. In a small missile it is practically impossible to achieve this physical separation and electrical separation is complex and difficult to achieve. For these reasons it will be found that for active radar seekers the pulse system is the most practical.

The most severe problem which the designer of any active radar homing system will have to face is the limited space available in the missile. Among other variables, the maximum range of a radar system is dependent upon transmitted power and antenna cross section. Since a twofold increase in range can only be achieved by increasing the transmitter peak power sixteen times, it is understandable why designers of airborne radars usually depend upon increase in antenna size to achieve additional range capability. In the case of the guided missile, particularly in the air-launched missile, a limit is soon reached on the size of the antenna that can be used. For these reasons the active radar guidance system is inherently a short range device.

The most obvious application of the active homing guidance system is against aircraft targets and in both the air-to-air and surface-to-air application it has an advantage over other systems which cannot be ignored;

namely, once the guidance system has obtained a lock-on, the missile system is capable of independently solving the fire control problem. This means that the traffic-handling capability of the system is far better than the command type guidance system and somewhat improved over the beam rider. In either application the traffic-handling capacity is, of course, not unlimited because a certain sequence of events must take place prior to launch of the missile. In the air-to-air case the target must first be acquired by the pilot optically or by means of the airborne intercept radar. The pilot must then fly the interceptor to the correct heading and range to assure missile radar lock-on. Upon launching the missile, the pilot is then free to break off the attack or launch additional missiles.

Although active seeker systems have possible application to air-to-surface and surface-to-surface systems, its use will be limited to those cases where there is some distinguishing characteristic of the target with respect to its background. Primarily, its use will be limited to isolated targets, such as ships at sea.

A semiactive homing guidance system is one wherein the receiver in the missile receives reflected energy from the target, the energy having been transmitted from a source other than the missile, as shown in Fig. 8.

The transmitter may be located at the launching station or at a point separated from the missile launching site. The transmitter may be located on the surface of the earth, as shown in the illustration or may be carried in an aircraft. As in the case of active homing guidance, the transmitted energy may be in the form of light, heat, radio, or sound waves. The main difference between the two systems is that the semiactive system is not independent of outside sources, since its guidance

intelligence is derived from energy transmitted from a point external to the missile.

In the case of the active radar homing system, it was stated that because of restrictions imposed by the size of the missile, such a guidance device was relatively short range. It is for this reason, among others, that semiactive radar guidance systems become operationally attractive. Taking for example a ground-to-air missile system, it should be apparent that since the transmitting portion of the system will be on the ground, the transmitter power and antenna cross section can be increased considerably over that possible in a missile. If, in fact, the power output can thereby be increased four-fold and the transmitting antenna area tenfold, then the theoretical maximum range will be increased 2.5 times over the active seeker case as shown by the simplified calculation,

$$\frac{R_s}{R_a} = \left(\frac{P_s G_T}{P_a G_R} \right)^{1/4} = 1.4 \left(\frac{G_T}{G_R} \right)^{1/4} = 1.4 \times 1.7. \quad (1)$$

Another advantage of semiactive homing from the overall system standpoint should not go unnoticed and that is its use in combination with other midcourse guidance means for terminal guidance to achieve comparatively long ranges with all the accuracy inherent in homing systems.

Since in many systems the target must be tracked by a fire control type radar in order to launch a missile of any kind, then this illumination can be used by the missile to home semiactively on the target. However, the missile does not need to home all the way from launching point to target, but may be controlled through a midcourse phase by command guidance or beam-rider guidance until it reaches a point close enough to the target to start homing on it.

As pointed out earlier in this paper, the accuracy of a command system or a beam-rider system deteriorates with range, but a homing-all-the-way system, whether active or semiactive is essentially short range as compared with them. The advantage of this combination guidance scheme is that it provides relatively long range while still retaining the accuracy inherent in a homing system. Of course, as always, we must pay a price for this gain and in this case the price we must pay is that the illuminating radar must remain trained on the target until intercept has been achieved, although any number of missiles can be launched against a single target, only one target can be handled at a time.

When semiactive homing is used for an air-to-air type missile, the missile is launched by a fighter or interceptor aircraft and the target is illuminated by the interceptor type radar carried by the attacking aircraft. The method of acquiring the target and launching the missile will be identical with those where an active missile homing system is used. However, with the active system the aircraft is free to break away to attack another target as soon as the missile is launched. This is not true for the semiactive case. The illuminating radar, as in the surface-to-air case, must be pointed at the target continuously until missile intercept is achieved. This is generally accepted as a liability of the semiactive system; however, a study of the parameters involved will show this to be a minor consideration. Due to the short missile flight time involved, the limited maneuverability of high-speed aircraft, and the ability of modern tracking radars to maintain tracking lock over wide angles and angle rates, it is possible with semiactive systems to execute the breakaway maneuver with practically the same speed as with an active system.

Another advantage of the semiactive system not apparent at first glance is that we

now have a system which automatically overcomes the basic difficulty previously encountered which is inherent to a CW type system, the antenna feed-through problem. As I said before, since the receiver and transmitter of a CW system must be on at the same time, the receiving and transmitting antennas must be separated physically or electrically. In the case of the active homing system, we found this difficult to achieve, but in the semiactive system we have a built-in physical separation of the antennas. This, therefore, makes it feasible to use CW radar techniques for a semiactive homing system where it is exceedingly difficult to do so for an active homing system.

6. COMBINATION GUIDANCE SYSTEM

I have just mentioned the possible combination of a semiactive homing system with a beam rider and have shown the advantages of such a system. In the practical case it will be found more often than not that a missile system will require a combination of guidance techniques in order to best meet the tactical requirement.

The missile guidance problem may be considered in three parts: launching guidance, midcourse guidance, and terminal guidance. It is in the surface-to-surface type missile that we will be most apt to find it necessary to use a combination of all parts because of the long ranges involved. When, for example, the midcourse phase is a hyperbolic or circular radio navigation system, such as described earlier in the text, the missile may well be launched at some distance from the guidance system. It will be necessary under these circumstances to provide some form of simple inertial guidance or command guidance to place the missile within the midcourse guidance pattern. If the warhead lethality pattern is such as to require extreme accuracy in the terminal phase of

such a missile flight, it will probably be necessary to add some form of terminal guidance to such a system. This terminal guidance could take the form of a passive homer or in the case of discrete radar targets, it might be an active radar homer.

In the case of surface-to-air missiles, one might use a homing-all-the-way system, a beam-rider system, or a command system if the ranges required were relatively short; however, long range, surface-to-air missiles will, in general, require combined midcourse and terminal systems. As pointed out earlier, the command and beam-rider system will lack the necessary accuracy at long range and will have to be augmented by some form of terminal homing which may be active,

semiactive, or passive to fit the tactical situation. The homing-all-the-way system is inherently short range and will have to be combined with a midcourse system to achieve the necessary range.

In the air-to-surface type missile, we have a similar situation. Short range missiles can probably achieve the necessary range and accuracy with a command, beam-rider, or homing system, but when a long range missile is called for we must go to a combined system to achieve the range and accuracy required. In the air-to-air application, the missiles will generally be in the short range class and a single form of guidance may be used. As stated earlier, beam-rider, active, semiactive, and passive homing may be used in this application.

REFERENCES

1. Locke, Arthur S., "Guidance," D. Van Nostrand and Company, Inc., New York, N. Y., 1955.

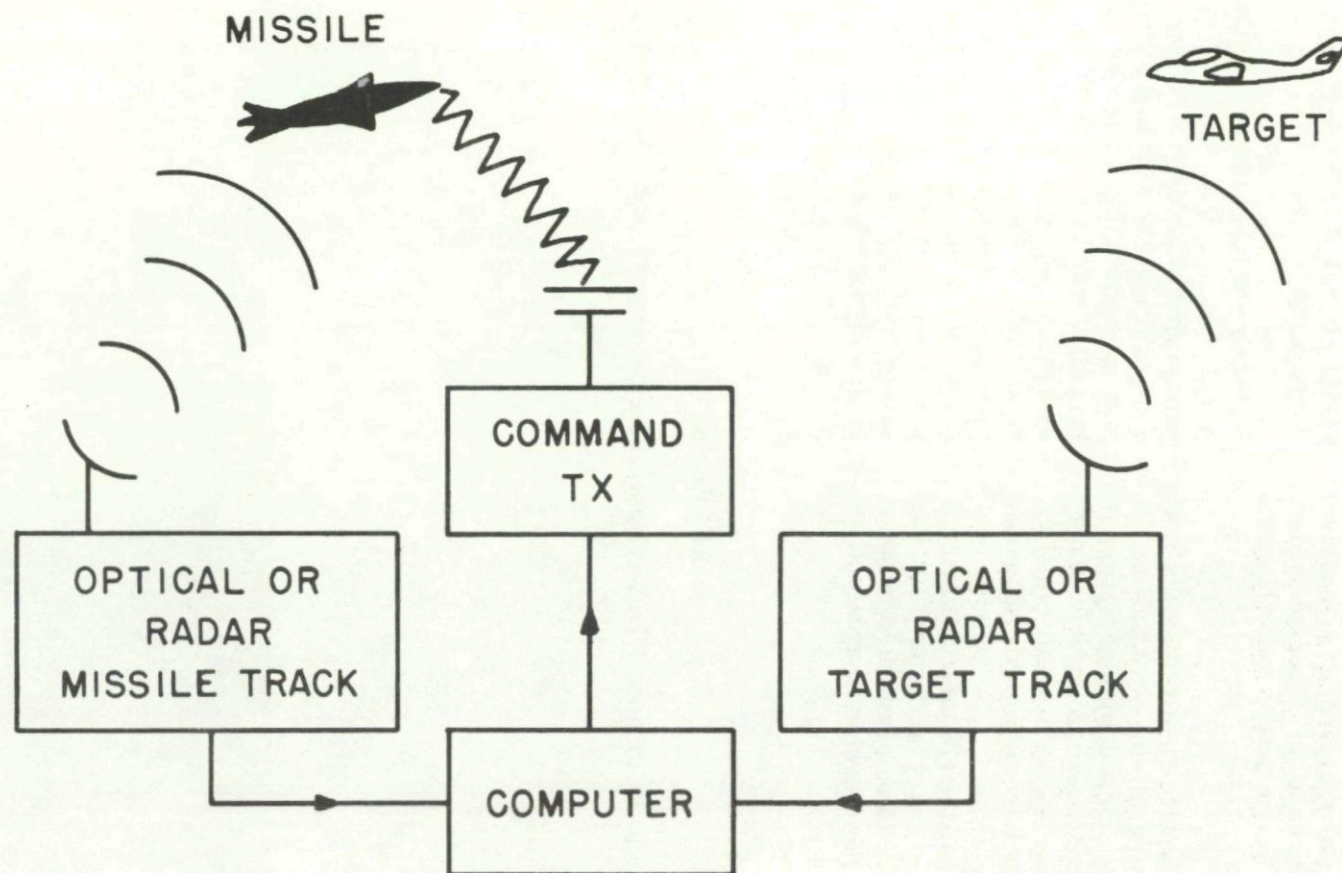


Fig. 1. Two-beam command system.

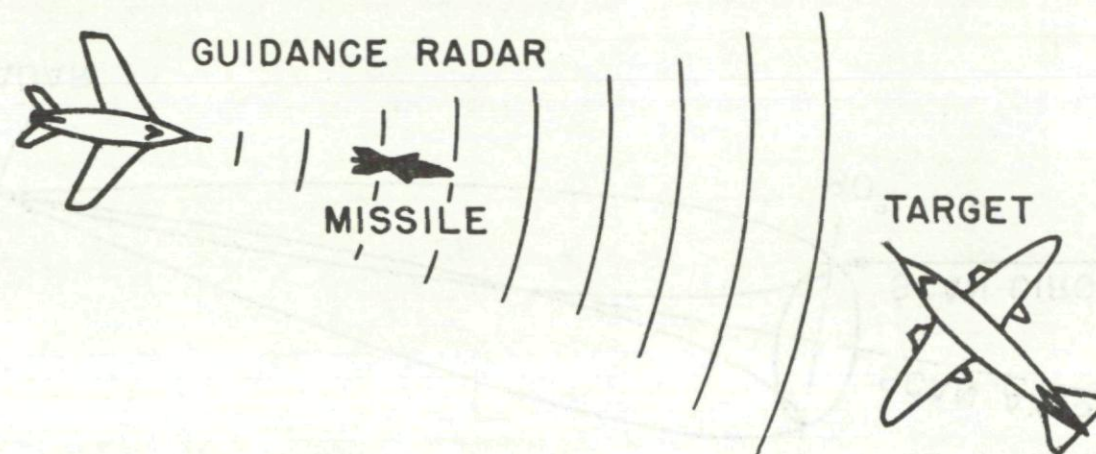


Fig. 2. Beam-rider system.

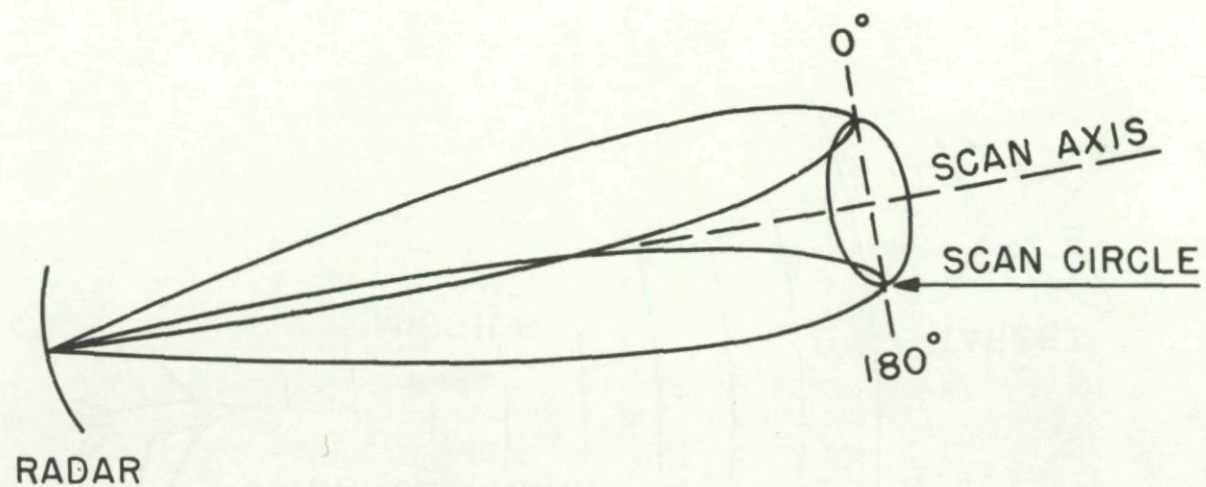


Fig. 3. Scanning radar system.

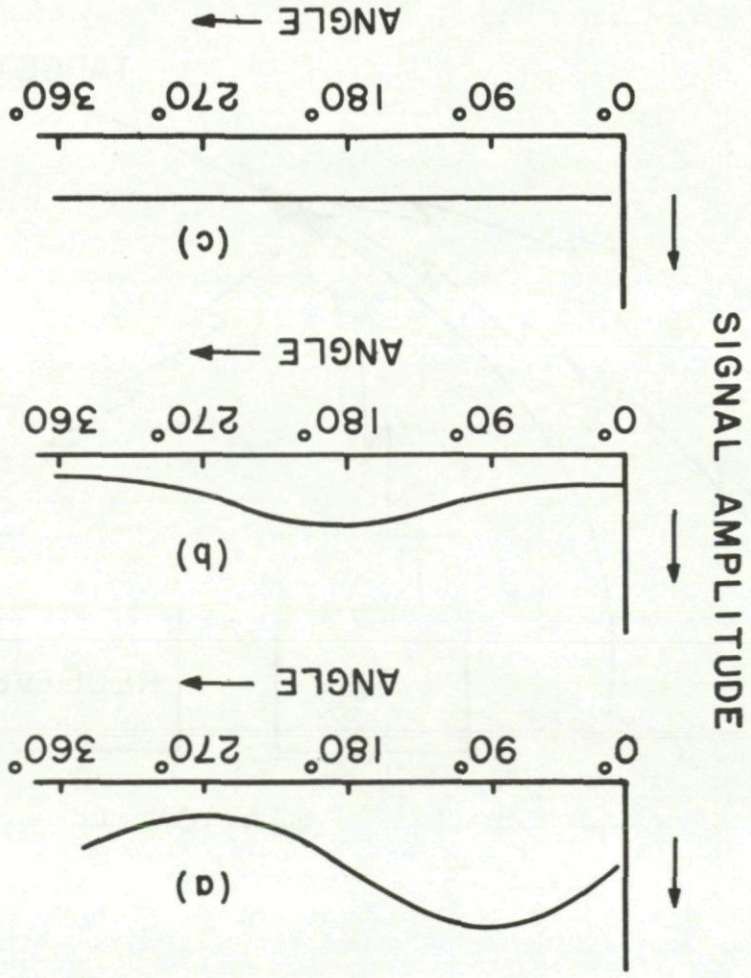
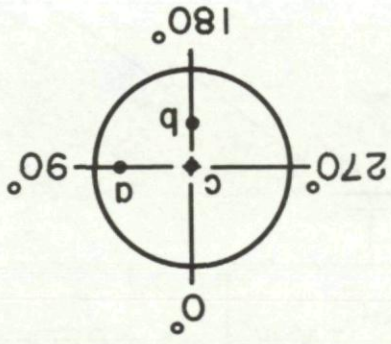


Fig. 4. Signal at output of scanning radar.

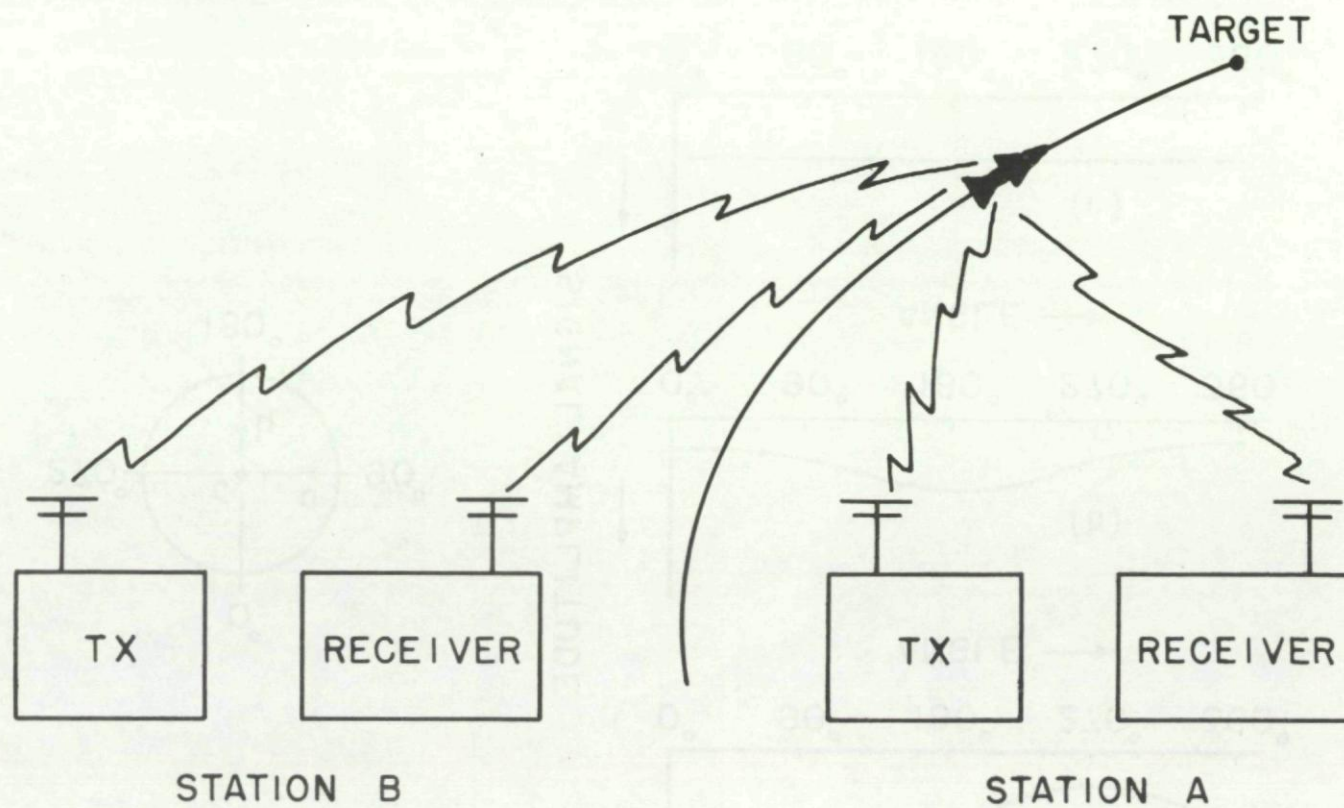


Fig. 5. Hyperbolic navigational system.

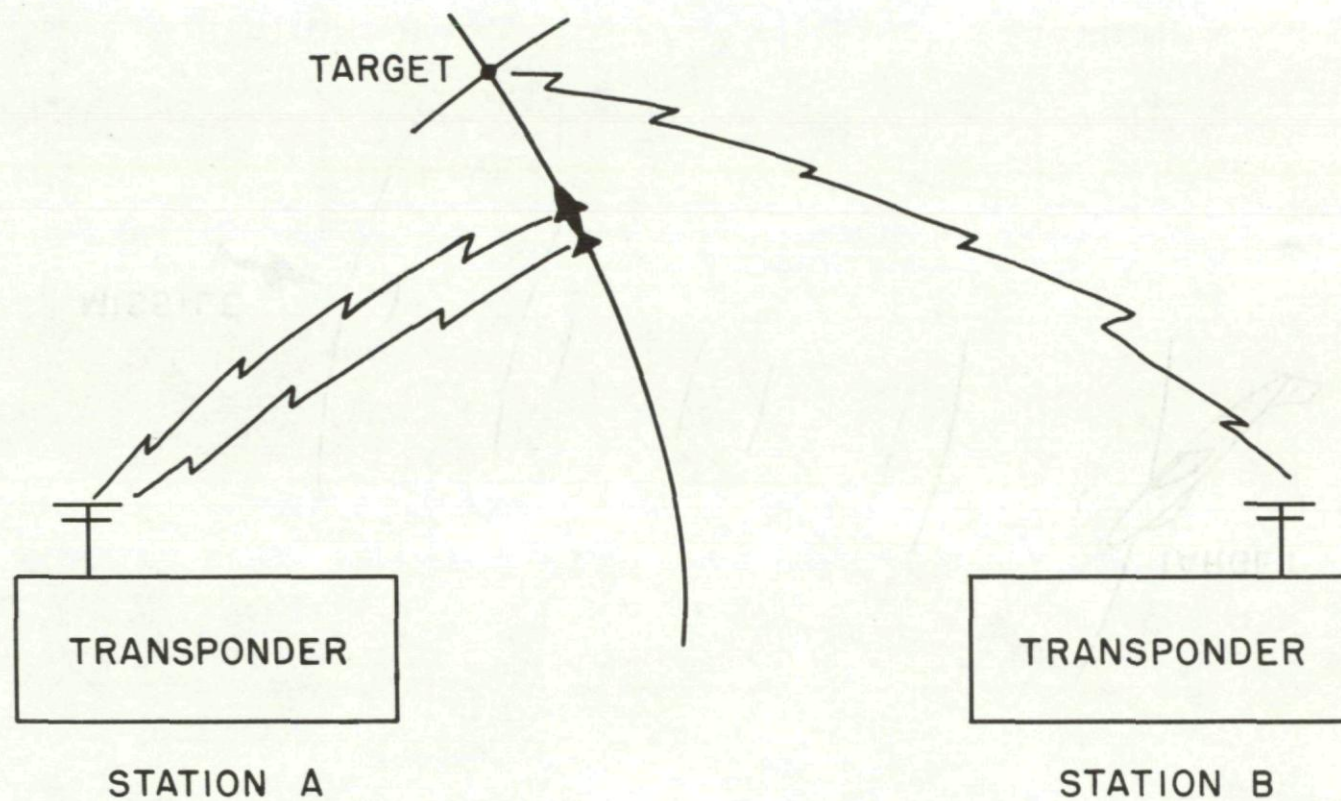


Fig. 6. Circular navigational system.



Fig. 7. Active homing guidance system.

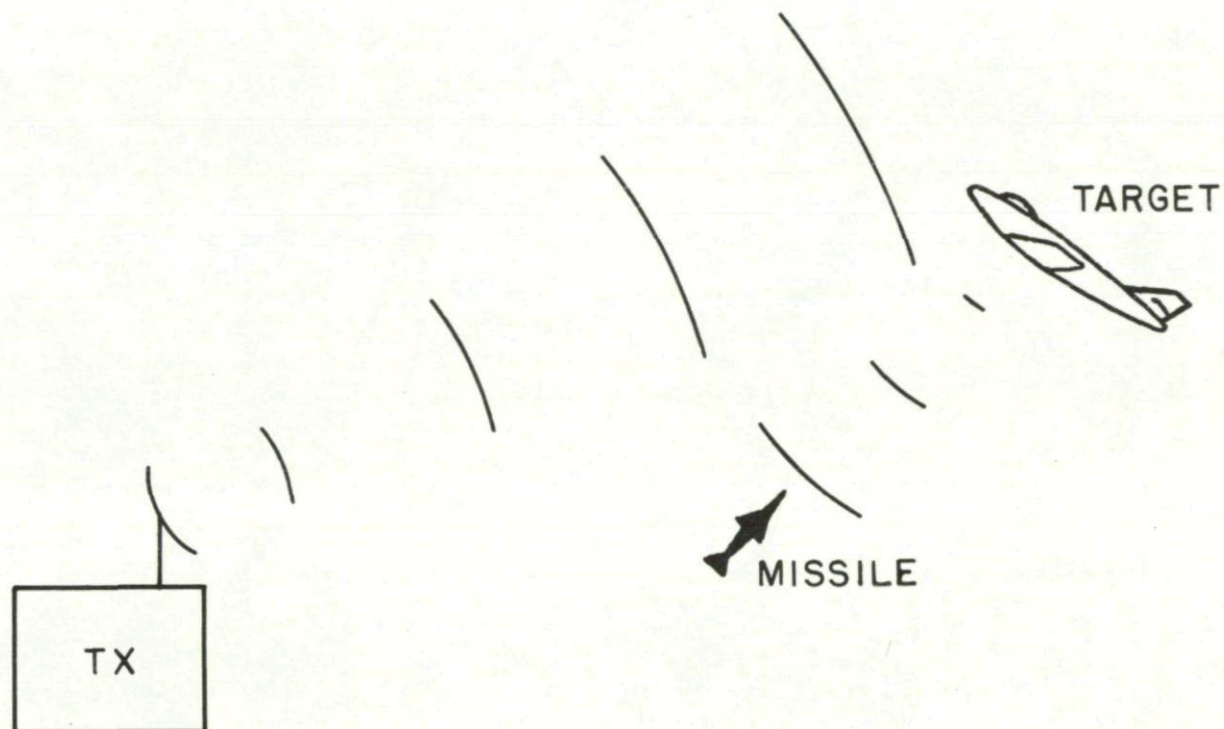


Fig. 8. Semiactive homing system.

CONSIDERATIONS IN THE CHOICE OF
A MISSILE GUIDANCE AND CONTROL SYSTEM
Robert William Mayer*

SUMMARY

This paper presents some of the major factors which should be considered in making a proper choice of a missile guidance and control system. The first part discusses the importance of thoroughly establishing requirements before making a choice. The second part enumerates the various criteria for judging a system against such requirements. In summary, the paper points out that there exists a great need for developing further techniques to provide better means of weighting the criteria in making a system choice.

SOMMAIRE

Cette note présente quelques uns des principaux facteurs qui devraient être considérés dans un choix correct du système de contrôle et de la gouverne d'un missile. La première partie traite de l'importance d'établir complètement les performances demandées avant de faire un choix. La deuxième partie énumère les différents critères permettant de juger un système sur de telles performances. En résumé, cette note montre qu'il existe dans ce domaine, un grand besoin de développer des techniques plus poussées en vue de fournir des meilleurs procédés, en faisant le choix d'un système, d'évaluer les critères.

1. INTRODUCTION

The objective of this paper is to present the major factors which should be considered in making a choice of a missile guidance and control system. I chose this objective because (1) in many cases, the problem of making the proper choice of system for a given application is equally as difficult as the problems of detailed synthesis and analysis of a system, and (2) time spent in making the proper decisions on choice of the system, in the beginning of a development, can result in savings in time and money in later phases of the program.

To meet this objective, I should like to stress the importance of thoroughly establishing the requirements of the system before

making a choice. I shall then discuss some of the more important criteria which should be considered in making such a system choice, and illustrate some of these with examples.

2. ESTABLISHING THE REQUIREMENTS
FOR THE GUIDANCE AND CONTROL
SYSTEM

It is extremely important that the requirements of a system be clearly established and understood before attempting to make a choice. The best designed system will be a poor one if it does not satisfy the requirements of the application. To insure that the correct system is chosen for the job requires consideration of factors described in the following paragraphs.

*General Electric Company, Philadelphia, Pennsylvania.

a. The Mission of the Complete Weapon System

In designing any system or subsystem which is a part of a complete weapon system it is impossible to specify in detail all the requirements of the subsystem; therefore the guidance and control system designers must have sufficient knowledge of the overall system. This knowledge allows them to exercise judgment in making a choice where specific requirements are not, or can not, be defined. For this reason they need to know the overall objectives of the weapon system, such as:

- (1) The nature of the target and tactical or strategic conditions under which it will be engaged.
- (2) The type of payload to be carried and its characteristics. (In some cases, payload choice should be concurrent with guidance choice.)
- (3) The vehicle from which the missile will be launched and its motions, if any.
- (4) General environmental conditions, i.e., where does the action take place?

b. Mission of the Guidance and Control System

If the designer has the knowledge of the mission of the overall system, he can then consider the more specific mission of the guidance and control system. This is defined by such specific factors as:

- (1) Range. How far must the missile be guided?
- (2) Altitude. At what height is the warhead activated?

(3) Accuracy. How close must it come to the target?

(4) Reliability. What percent failures can be allowed due to the guidance and control system?

These four factors define the basic objective of the system, i.e., what it must accomplish. They do not, however, show how the system accomplishes these objectives within the framework of the overall system. The definition of "how" is contained in a series of detailed specific requirements.

c. Specific Requirements

Definition of these specific requirements requires a knowledge of the overall system scheme and the interfaces which exist between the guidance and control system and other subsystems.

Such requirements can, of course, become quite detailed but, in general, should cover areas such as the following:

- (1) The trajectory scheme; defining where important functions take place, such as burnout, arming and fuzing, maneuvers, etc.
- (2) Power supplies available and their characteristics.
- (3) Propulsion characteristics.
- (4) Space and weight allocations, both in airframe and ground equipment.
- (5) Environmental conditions before and during flight.

d. Areas of Flexibility

Not all requirements of the guidance and control system should be fixed early in the design stage. Instead, some of the requirements of the guidance and control system should remain flexible over a given design range. This will provide more latitude in fixing design parameters in critical portions of other subsystems and at the same time provide the greatest possibility of a more optimum overall missile weapon system. In order to realize maximum gains from this flexibility it is necessary that the guidance and control system designers understand the limitations and problems of the other subsystems. They must understand, too, the interrelationships which exist between these subsystems and the guidance and control system. This thorough understanding can aid the complete integration of the guidance and control system with the overall system.

Some of the limitations to be considered in maintaining flexibility of requirements are:

(1) Limitations on the propulsion system. Very definite limitations exist in the repeatability of cutoff and difficulty in maintaining thrust and total impulse within close tolerances. For example, the guidance and control designer should consider if it is worthwhile, from an overall system standpoint, to have a complex thrust control in order to simplify the guidance scheme.

Perhaps elimination of thrust control would not cause a proportionate increase in guidance system complexity.

(2) Limitations of power supply. Power supply requirements should remain flexible until it has been established definitely whether or not it is best to build complex regulating supplies or provide increased complexity in the guidance circuitry to accept wider power supply variations.

(3) Limitations in airframe. A certain amount of flexibility in requirements should exist to allow for the best balance between guidance and control complexity and airframe complexity, in such areas as:

- (a) Backlash tolerances on control surfaces.
- (b) Tolerances on misalignment of airframe surfaces and sections.
- (c) Tolerances on thrust misalignment.
- (d) Tolerances in determination of aerodynamic coefficients.

The requirements for the guidance and control system define the design goals. It is important, especially during the preliminary design stage, that these be considered as goals and not as rigid requirements since deviations from these requirements may result in a more optimum overall system. As such, these design goals become a part of the criteria by which a choice of system should be made.

3. CRITERIA FOR CHOOSING A SYSTEM

In addition to the above requirements there are other factors of equal importance to be considered in the choice of a guidance and control system. I should like briefly to talk about them and the previously discussed requirements as criteria, and indicate some of the considerations which should take place to insure a constant balance of one criterion against the other when making a system choice. Fig. 1 is helpful in illustrating these factors.

a. Accuracy

When considering the requirement of accuracy as a criterion, it is important not to overemphasize it at the expense of other criteria. It is sometimes easy for this to happen since accuracy can be explicitly defined as a number, such as circular probable error, while, on the other hand, it is more difficult to attach a number to some of the other less tangible criteria, which will be discussed shortly.

Some flexibility should be permitted in the definition of accuracy. Rather than explicitly define accuracy as the probable error in, let us say, the X and Y coordinates, it may be advantageous to define it as circular probable error, thus giving more flexibility to the distribution of error between coordinates.

b. Reliability

The reliability requirement can be easily expressed as a number for use as a criterion. However, good methods are still not available for readily predicting whether or not a given reliability will, in fact, be achieved. Therefore, to obtain an explicit measure of how well the reliability criterion may be met by different choices of systems is presently a difficult job. Logically then, the consideration of reliability in making a choice between systems often must be carried out on a relative basis. Three major tools are available for predicting reliability: mathematics, experience, and information on the quantity and types of equipments.

(1) Mathematics. A figure for the reliability of a given system may be derived by predicting the reliability of the components within the system, based on either experience or test, and then combining the resulting values in the proper fashion to give an overall

reliability figure. This figure may then be compared with one obtained for a similar system, the component reliabilities of which were arrived at in the same manner. Reliability figures obtained in this manner are relative, rather than explicit, because of assumptions used for predicting reliability of components. However, they are sufficiently good for comparisons if like assumptions are used for component reliabilities.

(2) Experience. Experience plays a large part in assessment of relative reliability. If a system is made up of components which have already shown a high degree of reliability while operating in other systems, it can be expected that such a system will be the more reliable than one comprised of untried components. Of course, this is true only if the components are operated in the same environment. When assessing relative reliability in this way, it is sometimes helpful to list the components, indicating their previous uses and environments. By doing this, the system with the most "proven" components can be determined.

(3) Information on quantity and types of equipments. Information concerning the relative complexity of a system can be obtained by considering the numbers of like component parts involved, such as electronic tubes, resistors, capacitors, solenoid valves and relays. When doing so, however, thought must be given to the types of equipments involved and to the type of failure that may be incurred. In addition, consideration should be given to the probability of preventing this failure by a systematic maintenance inspection routine. Failure in mechanical components often can be predicted by visual inspection and gradual degradation of performance, whereas some electric components may fail completely without warning.

c. Accuracy vs. Reliability

A high accuracy system with poor reliability will result in few hits. In the same way, a low accuracy system with high reliability also results in few hits. Since greater accuracy generally calls for greater complexity and less reliability, it is important that both accuracy and reliability factors be balanced, as much as possible, to provide a system which will best insure the delivery of the payload to the target.

This balance, or compromise, could be made if we were able to draw curves for accuracy and reliability as a function of complexity and then combine the two types of curves to give a composite as in Fig. 1. Unfortunately, it is extremely difficult to do this, especially for the reliability criterion, since most of the reliability estimates are only relative, as was previously discussed. Keeping such a composite curve in mind however, is a help when making a system comparison.

d. Operational Characteristics

The ability of the system to meet the operational characteristics is equally as important as its achievement of other performance characteristics. Operational characteristics may be specified for guidance control designers as requirements or, in other cases, they may be presented only in a broad sense. When the latter is the case, the designers must insure that the equipment will meet their best estimates of the required operational characteristics. Specific areas which should be considered in making a choice of system are discussed in the following paragraphs.

(1) Ease of handling. The sizes and weights of missile and ground equipment pieces must be sufficiently small to insure

easy emplacement and handling in the preparation for launching. Moreover, the equipment must be designed for use in all of the environments that are to be encountered.

The answer to the question of whether to put more equipment in the missile or on the ground in the case of a surface-to-surface weapon is often influenced strongly by a desire for easy handling. For example, a tactical situation requiring the utmost ease of handling would cause one to choose a system having more of the guidance equipment in the missile, as in an inertial system, than on the ground as in a radio system.

(2) Alignment procedure. The ease of making initial settings and alignments is another important consideration since it has an effect on the type and quantity of equipment and manpower required to carry out the process. In addition, the rate of fire is influenced by the rapidity with which such adjustments can be made.

To insure that the best overall system is selected, it is important to balance complexity of on-board guidance equipment against complexity of ground alignment equipment. This is necessary because any alignment procedures requiring complex ground equipments may result in lower overall system reliability.

(3) Maintenance. Maintenance is an area which was neglected during some of the early phases of missile development. A choice between systems should lean heavily toward the one that does not require the services of systems designers to maintain proper operation. Skilled technicians alone should be able to maintain the selected system.

Of course, attention should be given to the location of equipments for ease of adjustment under the environmental conditions which will be experienced.

(4) Set-up time. Set-up time, which includes all operations necessary to locate, emplace, erect, and tune up equipment needed to launch the missile, is tied in closely with ease of handling. When a tactical situation involves set-up time, it can be an important factor in the choice of system.

(5) Rate of fire. The requirements for rate of fire are strongly influenced by the mission of the overall system. For example, a defending missile system for use against saturation attacks necessitates a maximum rate of fire. On the other hand, rate of fire requirements for strategic-type missiles may not be as stringent.

Rate of fire depends heavily on erection time at the launching pad, alignment time, out for maintenance, etc. For example, if a radar ground equipment is used for guidance, rate of fire may be limited by the amount of time the radar must track each missile during the course of its trajectory. This could be a strong factor in making a choice between a radio and an all-inertial type surface-to-surface missile system.

(6) Security from countermeasures. Security from countermeasures is one of the most important criteria on which a system choice should be made. In the case of a surface-to-surface missile, a consideration of this factor includes such questions as:

- (a) Can the radio guidance system be jammed?
- (b) Does the system involve a large ground guidance subsystem installation, which could easily be spotted from the air?

(c) Do the requirements of the guidance and control system cause the missile to be unnecessarily long and therefore easily spotted from the air?

(d) Are the alignment procedures so lengthy that the enemy has time to destroy the missile and associated equipment?

(e) Does the system require ground communication links which could easily be jammed or destroyed?

e. Requirements for Other Subsystems

The guidance and control system should be chosen such that the best possible system is obtained within the time scale allotted for development. To insure that the time scale will not be exceeded, care must be taken to make certain that the guidance and control system does not overcomplicate other subsystems. This subject already has been discussed somewhat under the subject of flexibility of requirements, but I would like to stress it further now.

In making a choice of system this criterion should be judged by asking such questions as:

(1) Does the choice of guidance and control system hinder or help the arming and fuzing functions? For example, a surface-to-air seeker system having no range information may depend on rate of change of line-of-sight or change in sign of range rate for fuzing. When compared with a system which supplies range information for fuzing this may impose a hardship on the arming and fuzing system.

(2) Does the choice of guidance and control system impose design hardships on the airframe in terms of space, arrangement, and aerodynamic control characteristics? For example, a surface-to-surface system where guidance equipment must be carried in the nose may result in difficult alignment procedures if the missile is erected vertically before launching.

Splitting of equipment into various compartments can also be troublesome by requiring more wires for interconnections and possibly rigid alignment specifications between compartments.

(3) Does the choice of guidance and control system cause complexity in the propulsion system by requiring thrust control, intermediate thrust stages, auxiliary control jets, etc.?

f. Growth Potential

The choice of guidance and control system should also include a consideration of the growth potential of the guidance and control system and the other missile subsystems. On examining the ability of systems to meet this criterion the designer should ask himself such questions as the following:

(1) Does the system provide for future incorporation of new components, presently under development, without change in basic system concept? For example, a guidance system might employ numerous components or correction devices in its early stages of development in order to meet accuracy requirements. However, by proper design, the system may be synthesized in such a way that, as the accuracy of basic guidance components increases, the correction devices may be merely eliminated without changing system concept.

(2) Will improvements in components make it possible to meet performance specifications which cannot presently be achieved without further system complication? If the answer is "yes," then it may be wise to choose the simple system and depend on component development to provide the improvement, rather than choose a more complex, less reliable system which will be burdensome in the future.

(3) Will the guidance and control system be quickly outmoded by improvements in other subsystems? It may be unwise to choose an overly complex guidance system to cope with deficiencies in other subsystems, if methods for eliminating these deficiencies are on the horizon. And if such methods are forthcoming, consideration should be given to choosing a guidance and control system that can easily be simplified as such developments become available.

g. Timing

The best system is useless if it is not available in time. Thus, it is of extreme importance to be sure that the selected system is sufficiently simple so that it can be developed in accordance with a schedule. In addition, care must be exercised to insure that critical components will be available in sufficient time. If the performance-improvement factor versus time is small, the system having the lesser performance and earlier availability should be favored.

h. Cost

The cost of a guidance and control system is of importance not only from the dollar standpoint, but also because it indicates the impact on the availability of manpower and facilities of designing, producing, and operating such a system. When considering costs, the following area should be explored to insure that the lowest cost, consistent with other requirements, is obtained:

(1) Manpower to develop critical items. If the system can be chosen to require a minimum of such items, cost will be lower.

(2) Ease of development testing. If a choice exists, it would be wise to choose the system most easily tested to determine performance. Consideration should be given to (a) the number of development models required and (b) the type of facilities required. A choice of a marginal equipment may require extensive environmental facilities for proof testing, while a choice of equipment that is already proven by other applications, having the same environment, would reduce such requirements.

(3) Production facilities. Is the design adoptable to mass production methods, such as printed circuit techniques?

(4) Cost to maintain and operate. Is an extensive staff of highly skilled technicians required to maintain the equipment? Is expensive special test equipment required? Will it stand up under long storage periods and will it withstand transportation environments without need for extensive replacements?

4. CONCLUSION

If requirements were perfectly written, criteria such as those presented, and others, would be an integral part of the statement of requirements; however, in many cases this

may not occur. It therefore becomes the duty of the guidance and control system designer to set up such criteria for judging the system.

To insure proper choice it is important that the procedure outlined, or one similar to it, be utilized. Often, because of a tight schedule a system may be selected without benefit of such analysis. If the choice is based on past experience with similar systems and requirements, it will probably be a good one. However, if the application is new, involving the use of newly developed techniques and components, an analysis such as that described should be made.

An approach such as has been outlined here is not an analytical one, it is more a matter of writing down the factors and of using judgment in the application of weighting factors to the various criteria. As such, some factors such as reliability may be more heavily weighted than is justified.

Thus, there exists today a great need for a better means of weighting the various criteria. The ideal approach would be to apply a weighting factor to each criterion and to add up in some fashion the total number of points for each system, the best choice being the one with the largest number. The big problem, however, lies in determining the weighting factors. Perhaps, as we all gain experience in this area of system engineering we can do better at applying such factors.

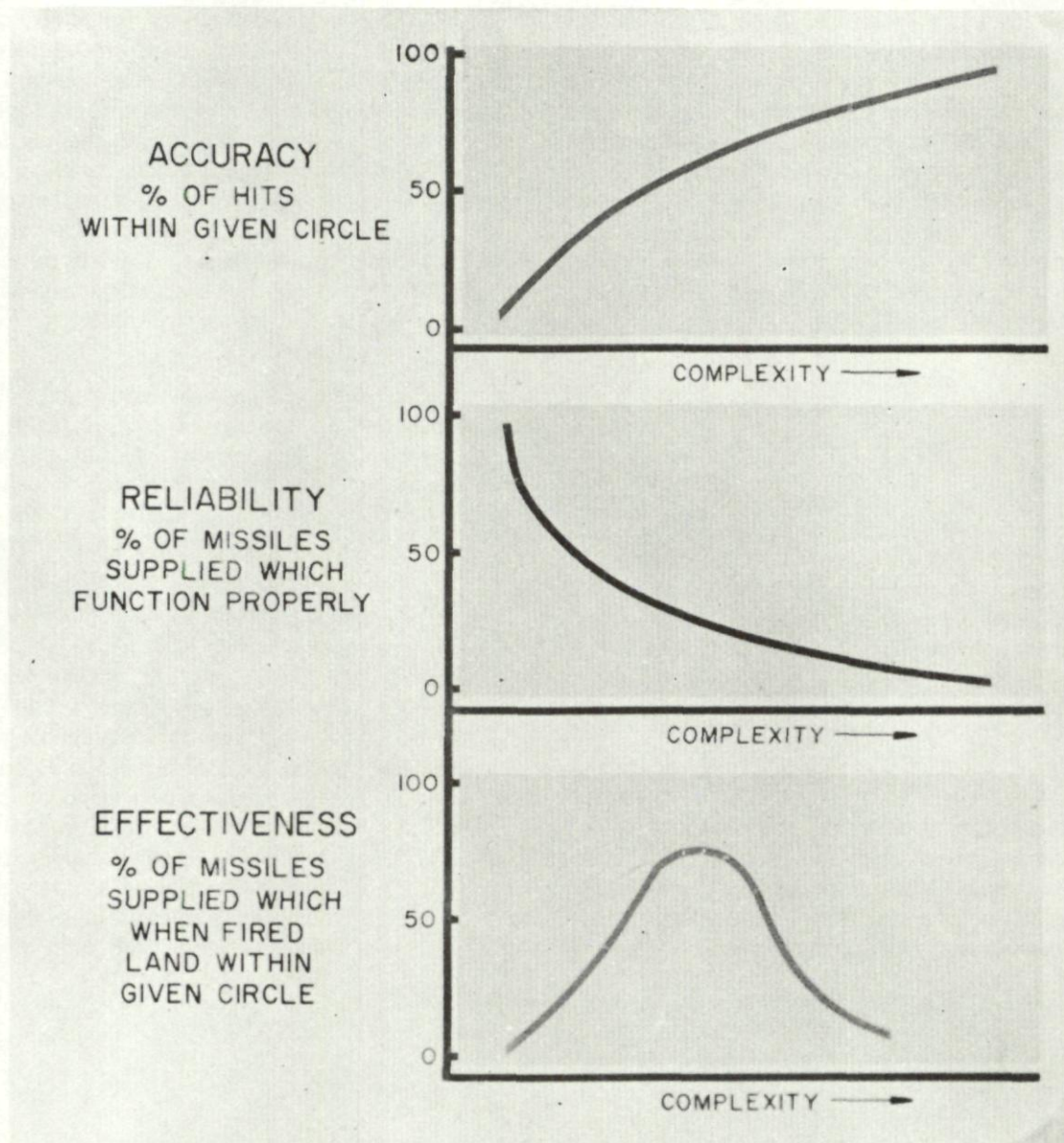


Fig. 1. Effects of complexity on system factors.

INERTIAL NAVIGATION

Norman F. Parker and Charles P. Greening*

SUMMARY

Inertial navigation is ideal for high-speed flight because of its inherent accuracy and freedom from the operational limitations of other navigation methods. It is an application of Newton's second law of motion, which implies that acceleration relative to inertial space can be detected without reference to external information. The basic elements of a physical system for inertial navigation are described. Accelerometers measure independent components of horizontal acceleration, while gyroscopes stabilize these accelerometers in a desired orientation. A computer finds position and velocity by integrating the acceleration components sensed in the vehicle, and also calculates orientation corrections for motion over the earth, rotation of the earth, and other factors. The source and propagation of certain errors in the system are mentioned, with emphasis on the initial misalignment of the accelerometers as an example. The navigation errors of the system which result from errors in the accelerometers and gyroscopes are stated.

SOMMAIRE

La navigation par inertie est idéale pour le vol à grande vitesse par son indépendance et son exactitude inhérente aux limitations opérationnelles des autres méthodes de navigation. C'est une application de la seconde loi de NEWTON qui implique que l'accélération relative par rapport à l'espace fixe, peut-être mesurée sans aucune référence à une information extérieure. Les éléments de base d'un système physique de navigation par inertie sont décrits. Les accéléromètres mesurent les composants indépendants de l'accélération horizontale; pendant que les gyroscopes stabilisent ces accéléromètres dans une orientation désirée. Une calculatrice trouve la position et la vitesse en intégrant les composants de l'accélération enregistrées dans le mobile et calcule également les corrections d'orientation dues au déplacement autour de la terre, à la rotation de la terre et aux autres facteurs. La source et la propagation de certaines erreurs dans le système sont mentionnées, par exemple, le cas d'un alignement initial incorrect des accéléromètres. Les erreurs de navigation du système qui résultent des erreurs dans les accéléromètres et dans les gyroscopes sont évaluées.

1. INTRODUCTION

Previous papers have indicated that a number of important changes are now taking place in the art and science of navigation. One such change is that inertial navigation, the determination of vehicle position and velocity by the measurement and integration

of vehicle accelerations, has become a practical reality and is assuming an important role in the navigation of high-speed aircraft. At the present time, the use of inertial navigation equipment is limited almost entirely to military applications. As a result, specific system details and performance are cloaked by security restrictions and are

*North American Aviation, Inc., Downey, California.

not generally known. However, evidence of the advancement which has taken place in this field is provided by the serious consideration currently being given to the use of inertial navigation equipment in commercial jet transport aircraft.

The basic principles of inertial navigation have been known for many years. In fact, the physical laws were formulated in the time of Galileo and Newton, and some of the more subtle effects were recognized by Jean Foucault, Max Schuler, and others. However, only recently, with the advent of high-speed aircraft, have the shortcomings of conventional navigation schemes become sufficiently important to justify the intense development effort required to make inertial navigation practical.

Dead reckoning techniques, based on air-speed measurements and compass direction information, are marginal as to accuracy. This is particularly true at high altitudes where jet streams may be encountered, near the earth's poles where magnetic compass accuracy is poor, and in flights in which considerable distance must be traveled between known landmarks or other checkpoints.

Radio and radar navigation techniques are very well suited for many applications, especially for guidance near the origin and destination of the flight. They depend on the receipt of signals from the ground and, hence, are of limited use in military applications. They are also of restricted use over large oceans, and are affected by magnetic storms.

Celestial navigation, based on measurements of the angular positions of stars relative to the local gravitational vertical, is adequate for navigation of ships and low-speed aircraft. However, as aircraft speed increases, this method becomes unreliable due to the inaccuracy involved in establishing the local vertical. The direction of the local

vertical cannot be determined accurately by conventional techniques because accelerations of the vehicle cannot be distinguished from the acceleration due to gravity.

An indication of the magnitude of error involved is given by the following example: Consider an aircraft on a course which a pilot considers to be straight but which is actually curved with a radius of curvature of 1000 nautical miles (1830 kilometers). At Mach 1, the centrifugal force associated with this very slight curve would introduce an error of approximately 17 nautical miles (31 kilometers).

Inertial navigation, on the other hand, is ideally suited for high-speed flight. The signal to be measured, namely acceleration, is large and the effect of the time buildup of error which occurs in inertial systems is reduced by the short flight times associated with high-speed aircraft. Although inertial navigation is also well suited for many applications in low-speed vehicles, it was the severe navigational requirements associated with high-speed flight that provided the impetus necessary to bring inertial navigation to its present state of development.

2. PHYSICAL BASIS OF INERTIAL NAVIGATION

Inertial navigation as used in this paper means navigation with respect to space. It is accomplished by mechanization of Newton's second law of motion which states that force is equal to the time rate of change of momentum. While position and uniform velocity are not directly detectable without reference to external information, it is apparent from a consideration of Newton's second law that acceleration is directly detectable. If a mass is suspended (Fig. 1) in such a way that it will be deflected from a neutral position by any force acting upon it,

and so that a restoring force arises in the presence of the deflection, this force or deflection can be used as a measure of acceleration with respect to space.

If the initial position and velocity of the accelerometer are known, position and velocity at any later time can be computed by detecting all accelerations and integrating with respect to time - once to obtain velocity and twice to obtain position.

3. MECHANIZATION

The simple accelerometer (Fig. 1) is so arranged that it is sensitive only to accelerations along one axis, called the sensitive axis of the instrument. If two such accelerometers are mounted with their sensitive axes at right angles to each other (Fig. 2), they can be used to measure any arbitrary acceleration in the plane defined by the two axes. If we consider an acceleration as indicated by the heavy arrow in a direction somewhat east of north, the north-south accelerometer will detect one component of the acceleration and the east-west accelerometer the orthogonal component. These acceleration components can be integrated separately and interpreted as distances north or south and east or west of the assumed starting point.

It is apparent that there is a need for stability of accelerometer orientation about all axes. If the platform upon which the accelerometers are mounted is misaligned in azimuth, as indicated (Fig. 3), an acceleration in the northerly direction which should produce a reading only on the north-south accelerometer will also be detected to a slight degree by the east-west accelerometer. The integrated output in this case will be slightly in error in both directions since the true northerly distance will be slightly greater than that indicated and the true east-west distance will actually be zero, although a value will be indicated.

It is also necessary to prevent the accelerometer platform from tilting about a horizontal axis (Fig. 3) because there is no way for any acceleration sensing device to distinguish between the acceleration due to gravity and an acceleration due to change in velocity.

The accelerometer platform can be stabilized, even in a moving vehicle, by mounting the platform in a gimbal system and providing gyroscopes which have the property of maintaining a fixed direction in space (Fig. 4). One possible arrangement involves the use of three separate gyroscopes, each of which has a single axis of free motion perpendicular to its spin axis. Each gyro will then provide stability about an axis perpendicular to the free axis and the spin axis (Fig. 5). A torque acting about a stabilized axis will be opposed by the gyro, which will precess about its free axis. The angle of precession can be used as a signal to counteract the torque by energizing servomotors in the stable element gimbal system.

If we expand the picture to include the effects of curvature of the earth and rotation of the earth, we find that corrections will have to be made to the platform orientation. Let us first consider the effects of motion over the surface of a spherical earth (Fig. 6). Because the stabilizing gyros maintain their direction in a space coordinate system, motion of the system over the surface of the earth will result in an apparent tilt of the platform relative to the earth. This tilt will introduce a component of gravity into one or both of the accelerometers, unless the accelerometer platform is driven through an appropriate angle.

Even though the system may be stationary on the surface of the earth, it will develop an apparent misorientation after a short time of operation due to the rotation of the earth.

That is, the platform will appear to rotate from east to west (like the stars) about an axis parallel to the axis of rotation of the earth.

To prevent the introduction of gravity effects into the accelerometers, it is necessary to correct the orientation of the platform for both of the above effects continuously.

The correction for motion over the earth requires knowledge of the angular motion of the system. The radius of the earth is known and the distance traveled is available from the accelerometers themselves. If the distance traveled is divided by the radius of the earth plus the altitude above the surface, the resulting angle is precisely the angle through which the platform must be tilted to maintain the proper orientation. The output of the computer which performs this division is applied to the platform controls in such a way as to keep the accelerometers always properly oriented.

Similarly, the orientation can be corrected for earth rotation by supplying a signal which is a function of time and latitude of the system (Fig. 7). It is apparent that an earth-stabilized platform will experience a component of the earth's rotation about its north-south axis, given by the cosine of the latitude, and a component about its azimuth axis, given by the sine of the latitude. The computer will compute the present latitude from the initial latitude and the second integral of the north-south accelerometer output. It will then resolve the constant earth rate into north-south and azimuth components, and send the portions to the appropriate platform drives.

In any but the crudest applications, accuracy requirements make it necessary to take into account secondary effects due to the oblateness of the earth and due to the angular

motion of any earth-fixed coordinate system (the well-known Coriolis effect). These effects require additional elements in the computer but introduce no new principles.

4. SYSTEM OPERATION

The block diagram (Fig. 8) shows the essential elements of an inertial navigation system. For simplicity, a single-axis system is shown. Its operation can best be illustrated by considering the effect of acceleration of the vehicle in which the system is located. An acceleration to the right, for example, will be detected by the accelerometer. The output of the accelerometer will be doubly integrated, and the resulting distance information displayed for the operator. In addition, this information may be used, after any necessary computation, to provide angular correction information to the stabilized platform. The stabilizing gyro will resist any inadvertent torque tending to tilt the platform and will provide signals to the gimbal servomotor to counteract the disturbance torque and maintain the platform in its proper orientation.

When the acceleration ceases, the accelerometer output drops to zero. The first integral remains constant, and the second integral continues to increase at a uniform rate, providing correct distance information and platform rotating signals. If the carrying vehicle slows and stops, the accelerometer senses the negative acceleration, causing the distance indicator to cease turning. Any increase or decrease in speed due to winds or control forces will be detected and integrated in the same way.

5. ERRORS

Any mechanical or electrical system can be expected to depart somewhat from perfect response to input signals. At first glance it

appears that the performance of the inertial instruments (that is, the accelerometers and gyros) will have to be surpassingly good in order to provide accurate position information over a period of hours. It might be expected, for example, that a fixed platform tilt would cause an error to build up as the square of the time. ($E = 1/2 \sin \phi g t^2$, where E is error, ϕ is tilt, g is gravitational acceleration, and t is time.) Similarly, gyro drift rate might be expected to result in an error that increases as the cube of the time. However, while it is true that the requirements on these instruments are severe, the curvature of the earth, together with proper design of the system, results in a limiting effect on system errors. This effect can best be illustrated by considering an example.

A typical source of error in an inertial navigation system is an initial misalignment or tilt of the platform. If we examine qualitatively the effects of a tilt on a single axis platform which is stationary, as in a laboratory or a standing aircraft, we can get a picture of error propagation uncomplicated by aircraft accelerations and by interactions between axes of stabilization.

Let us assume (Fig. 9) that the system is started and aligned, with an initial incorrect tilt, ϕ . The accelerometer registers an acceleration of $g \sin \phi$, which is integrated and interpreted as a motion to the left. Hence, the computer sends a signal to the platform tilting it to the left through an angle corresponding to the spurious travel in that direction over the curved earth.

At a later time (Fig. 10), the tilt to the left has equaled the misalignment angle, ϕ , and the platform is now level. There is now no input to the accelerometer. However, the first integral is still positive and the second integral is still increasing, causing a further rotation in the same direction, resulting in a

sensed gravity component opposite to the original one (Fig. 11). After a time the first integral is brought to zero and the platform comes to rest, this time with a tilt of $-\phi$.

The reverse of the original situation now exists, and the platform is tilted back toward the original orientation (Fig. 12) and continues to oscillate in this fashion.

This illustration (Fig. 13) shows graphically the acceleration, velocity, position, and platform angle as compared with the proper values for these numbers as a function of time in the situation we have just described. We have seen how the accelerometer detects a \bar{g} component which decreases through zero to the opposite polarity and then returns to the original value. The effect upon the first and second integrals is shown in the velocity and distance curves. The platform tilt, which is directly dependent upon sensed distance, is also shown.

The result of this behavior is to provide a bound to the position errors resulting from an initial error. If the tilt error were propagated in the usual fashion, as $1/2 (g \sin \phi) t^2$, an unbounded position error would result. However, because the earth is not flat and because terms related to the curvature are fed back into the platform controls, the error is limited to a value which oscillates within narrow, fixed bounds. The period of this oscillation, for a properly designed system is determined by the values of the earth's radius and "g," and is 84 minutes. It is, in other words, a Schuler tuned system. It is this phenomenon which makes inertial navigation practicable.

A null error (that is, an error in the location of the zero point), or a scale factor error in the accelerometer, will result in somewhat similar oscillating error curves (Fig. 14). The phase relationships and the

size of the mean error varies from one case to another, but the frequency and general form of the curves is similar. A drift in the levelling gyros introduces an error term which increases linearly with time, in addition to an oscillatory component.

For long navigation times, dynamical coupling among parts of the system must be considered. Also, error sources can usually be better described statistically than as constants. The net result of these two considerations is to reduce the long term buildup of error from that which might be expected on the basis of constant independent error sources.

6. COMPONENT REQUIREMENTS

The critical elements of inertial navigation systems are the accelerometers and gyroscopes. The present state of the computer art, both analog and digital, makes it possible to meet the accuracy and capacity requirements by careful design.

7. CONCLUSION

We have seen that it is theoretically possible to develop a navigational device which depends only upon its initial condition (position and velocity), and its accelerations relative to space in order to provide continuous position information. The advantages of such a system of navigation are numerous. It is independent of wind velocity and visibility conditions which often render more conventional navigation methods inaccurate or unusable. It is equally effective over land or water, during night or day, and places no restrictions on the course of the vehicle in which it is carried. Furthermore, since no external information is required after system activation, the system cannot be prevented from operating by electrical or magnetic fields of artificial or natural origin.

It is apparent that a system having these characteristics will find widespread applications and will have a profound influence on the military and commercial transportation problems of the future.

REFERENCES

1. Slater, J. M., "Choice of Coordinate Systems in Inertial Navigation," Navigation, Volume 5, No. 2, Journal of the Institute of Navigation, June 1956.
2. Slater, J. M., and Duncan, A. B., "Inertial Navigation," Aeronautical Engineering Review, Volume 15, No. 1, January, 1956, Institute of the Aeronautical Sciences.

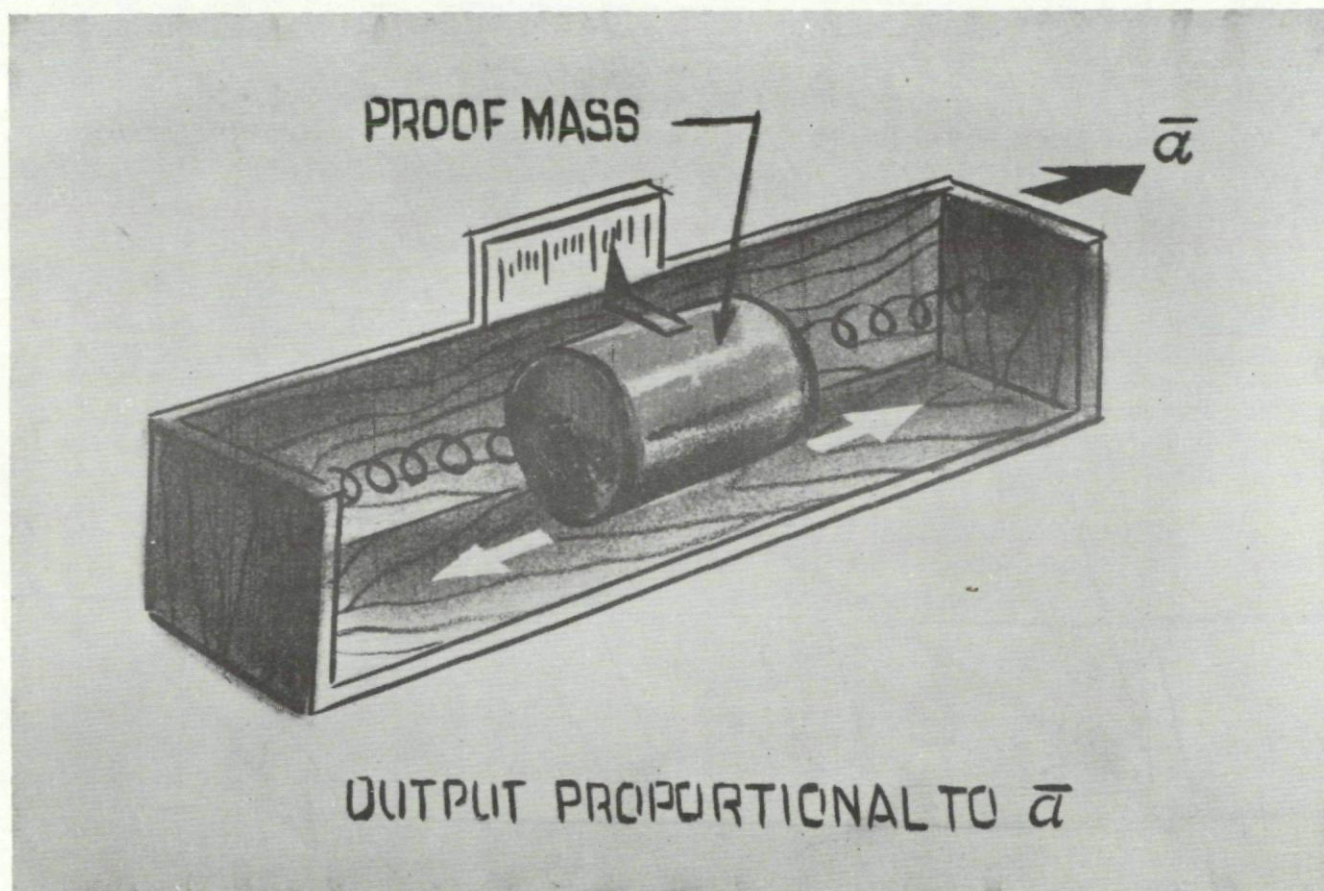


Fig. 1. Simple accelerometer.

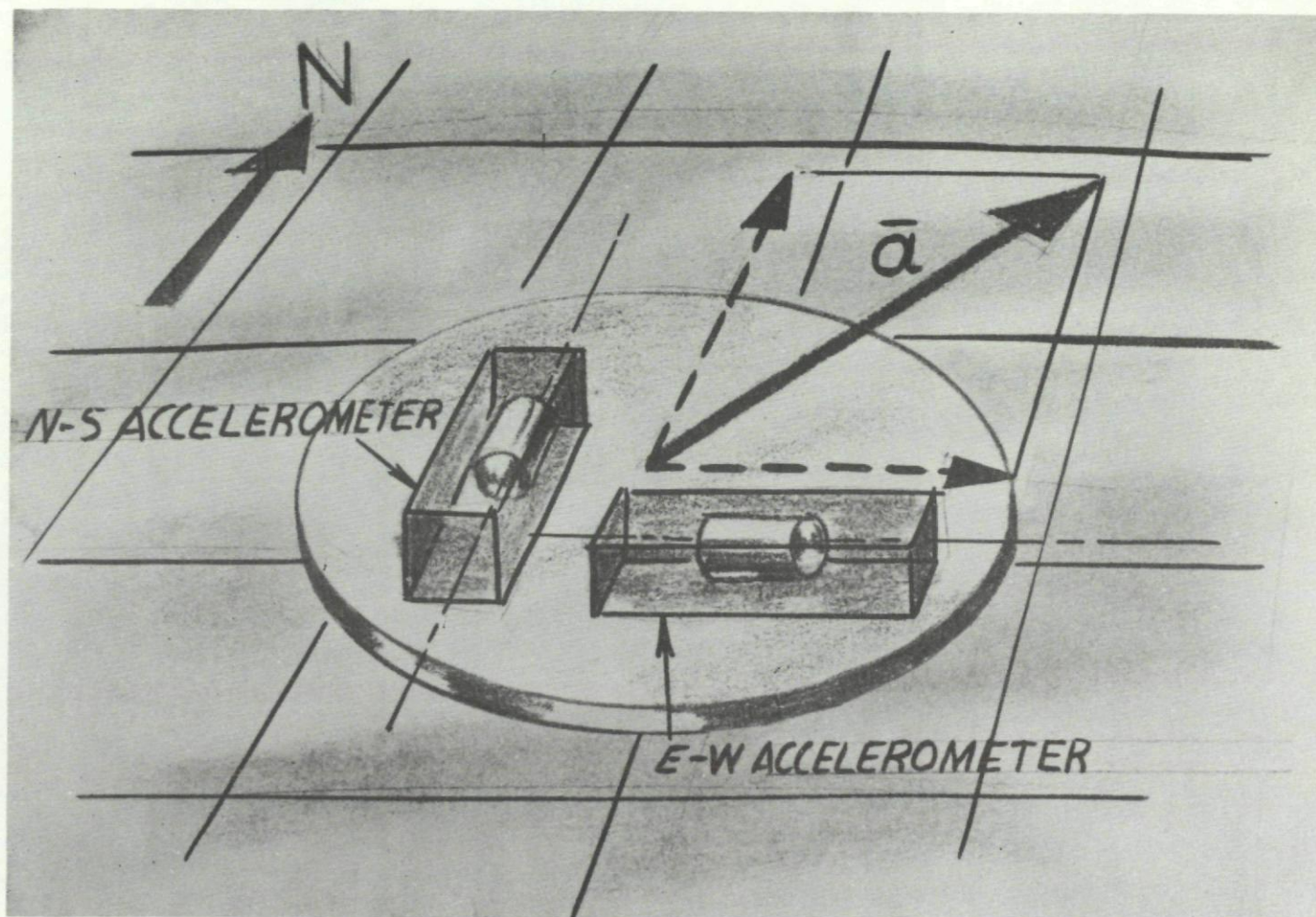


Fig. 2. Resolution of accelerations.

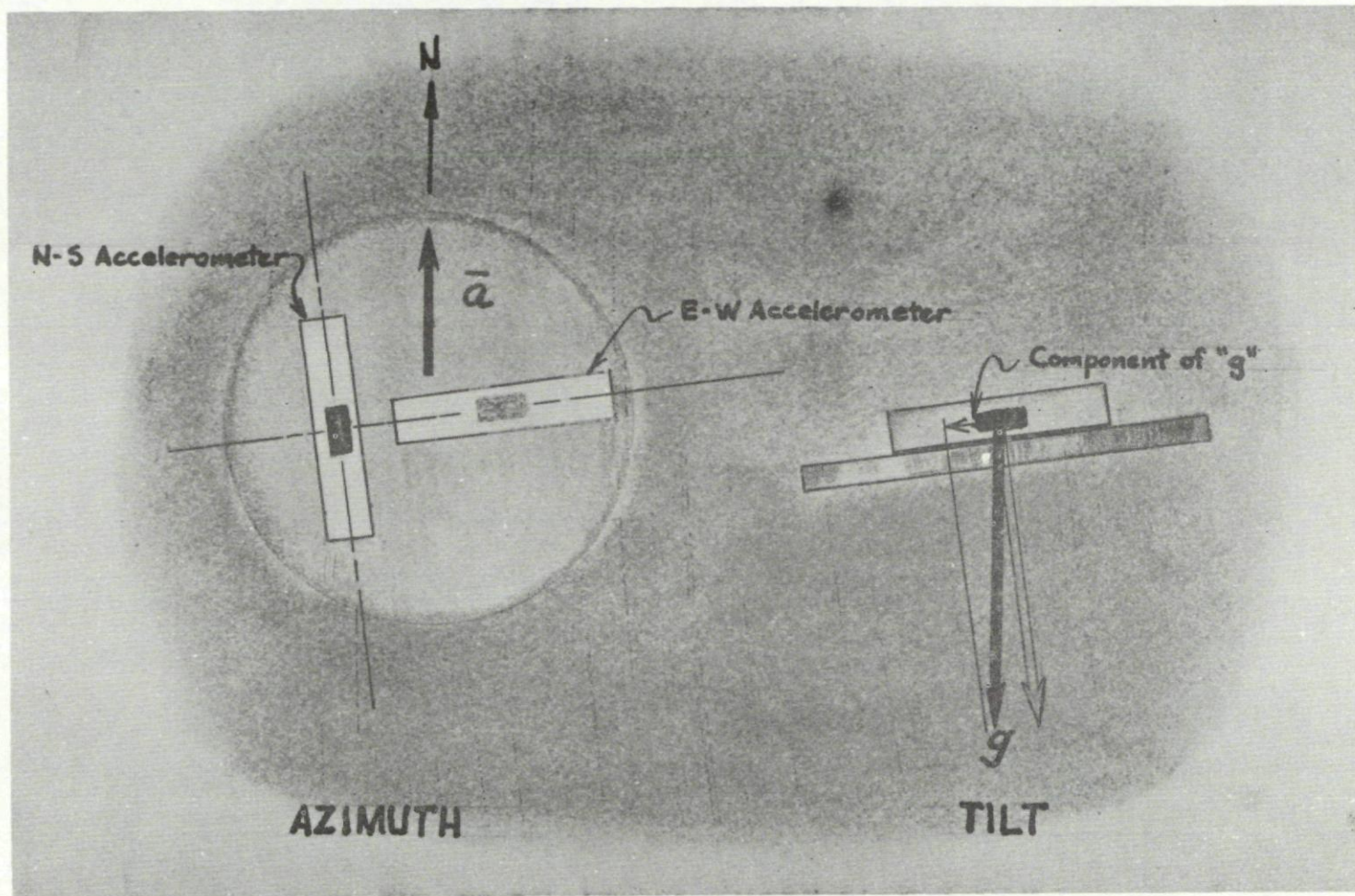


Fig. 3. Orientation of requirements.

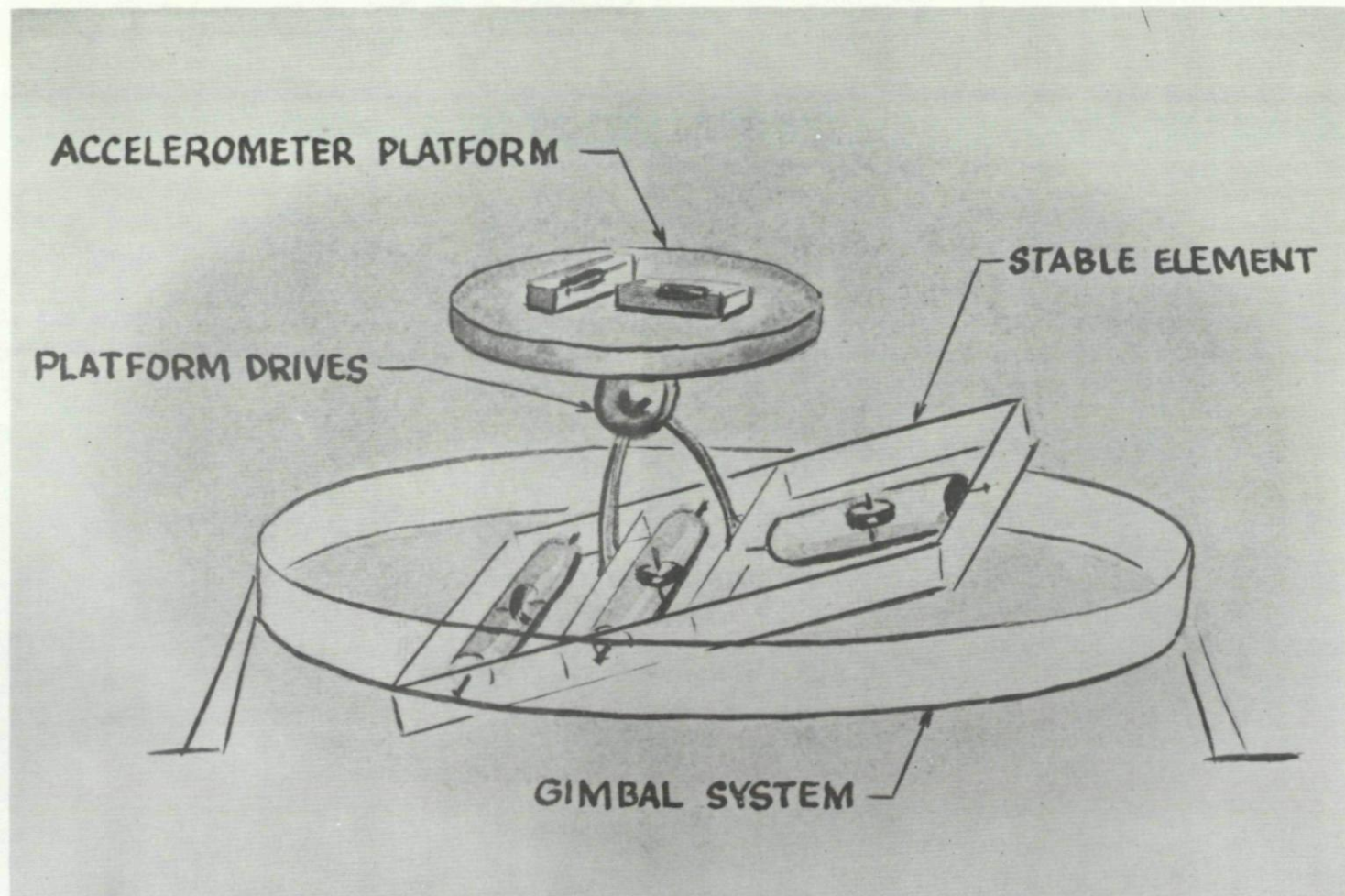


Fig. 4. Stabilization of accelerometer platform.

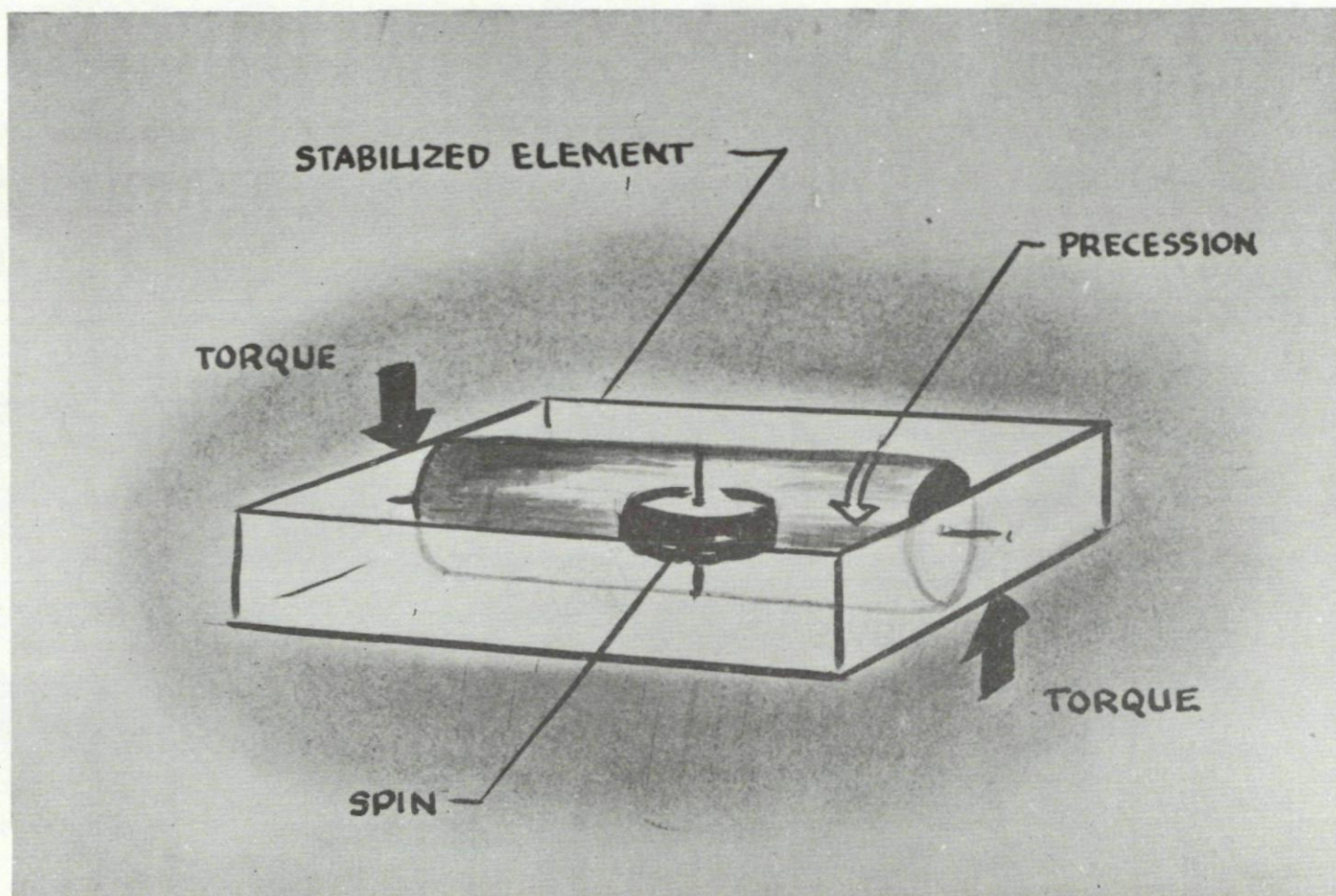


Fig. 5. Single axis gyro stabilization.

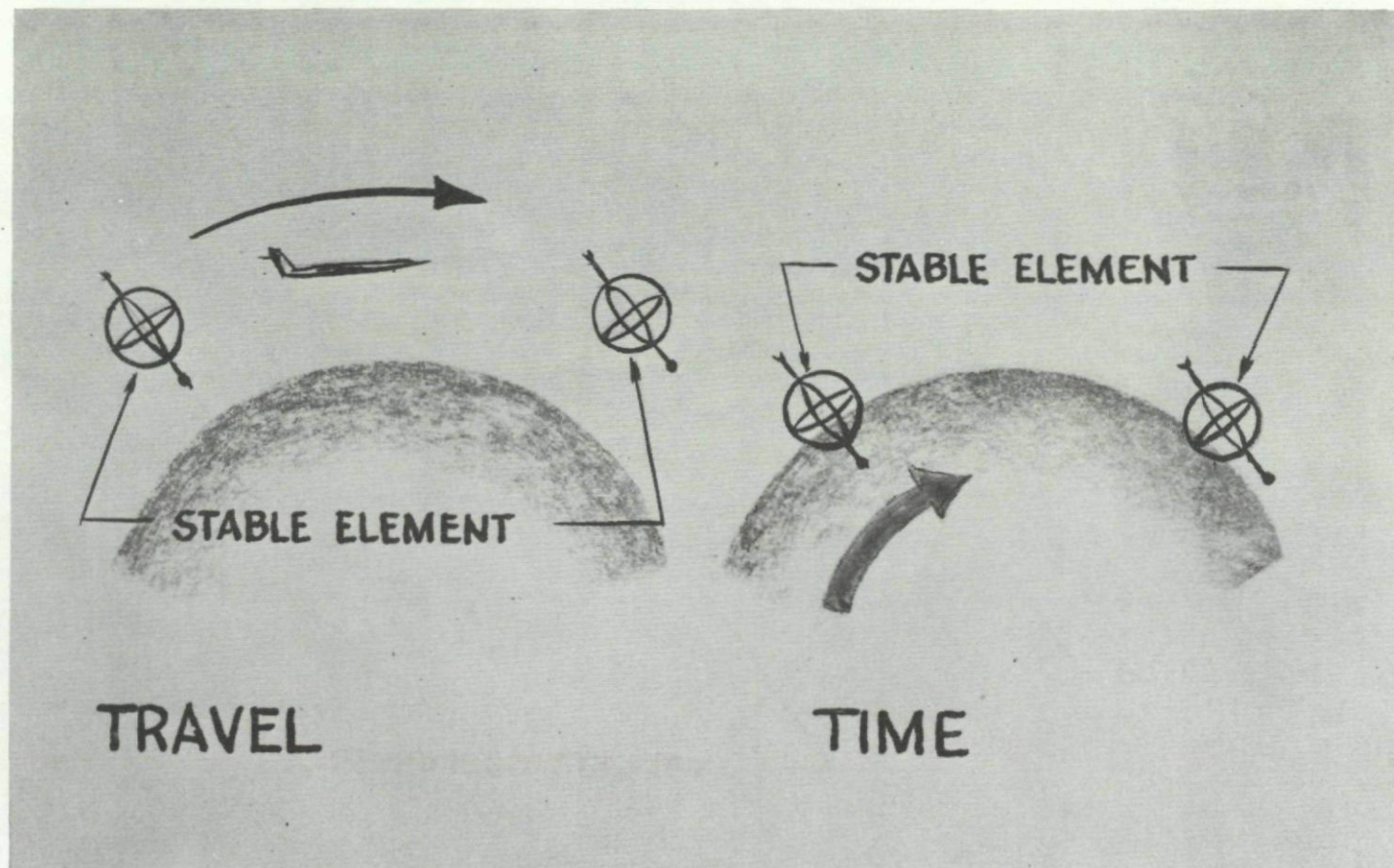


Fig. 6. Change in orientation.

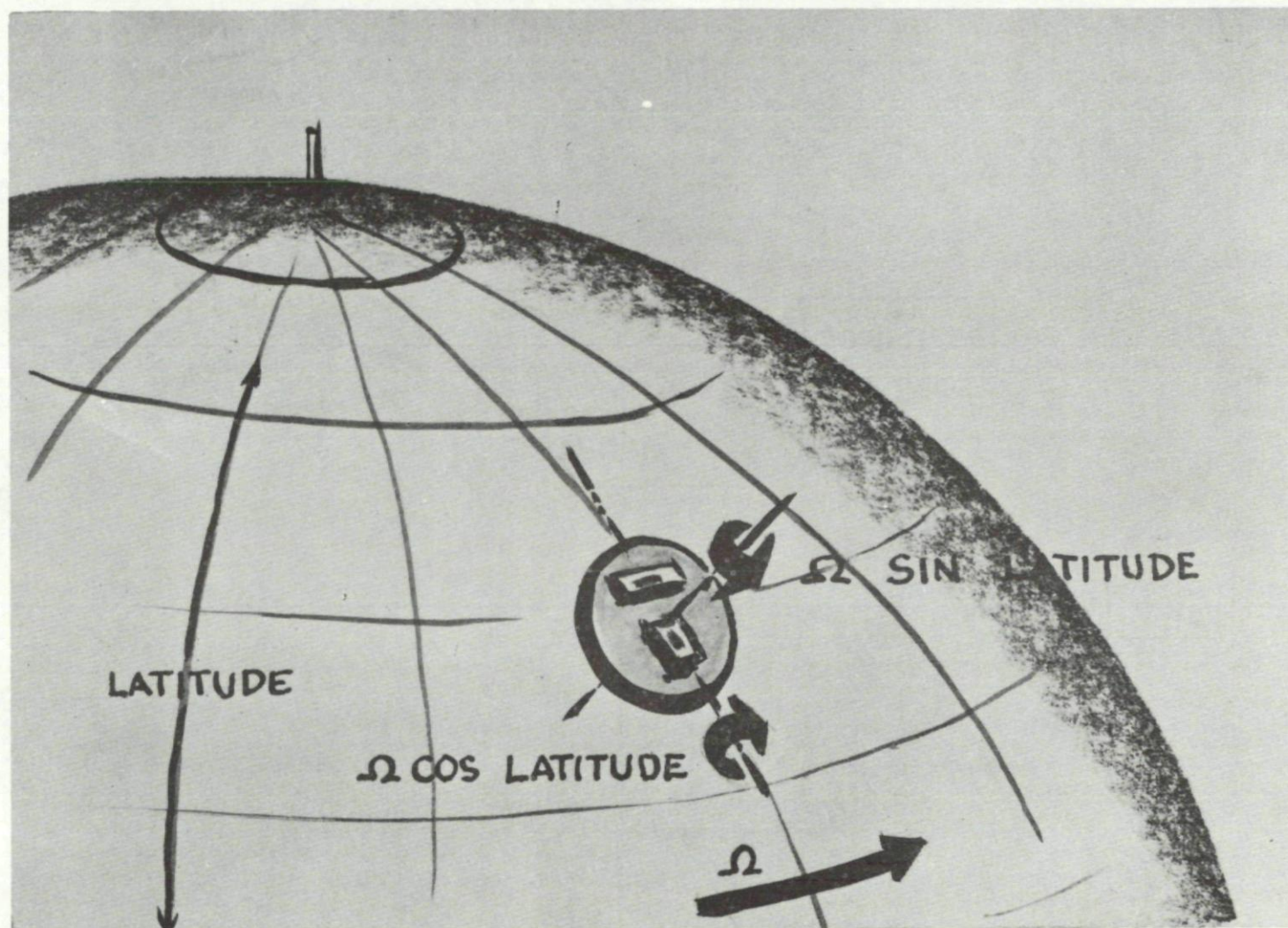


Fig. 7. Components of earth rotation.

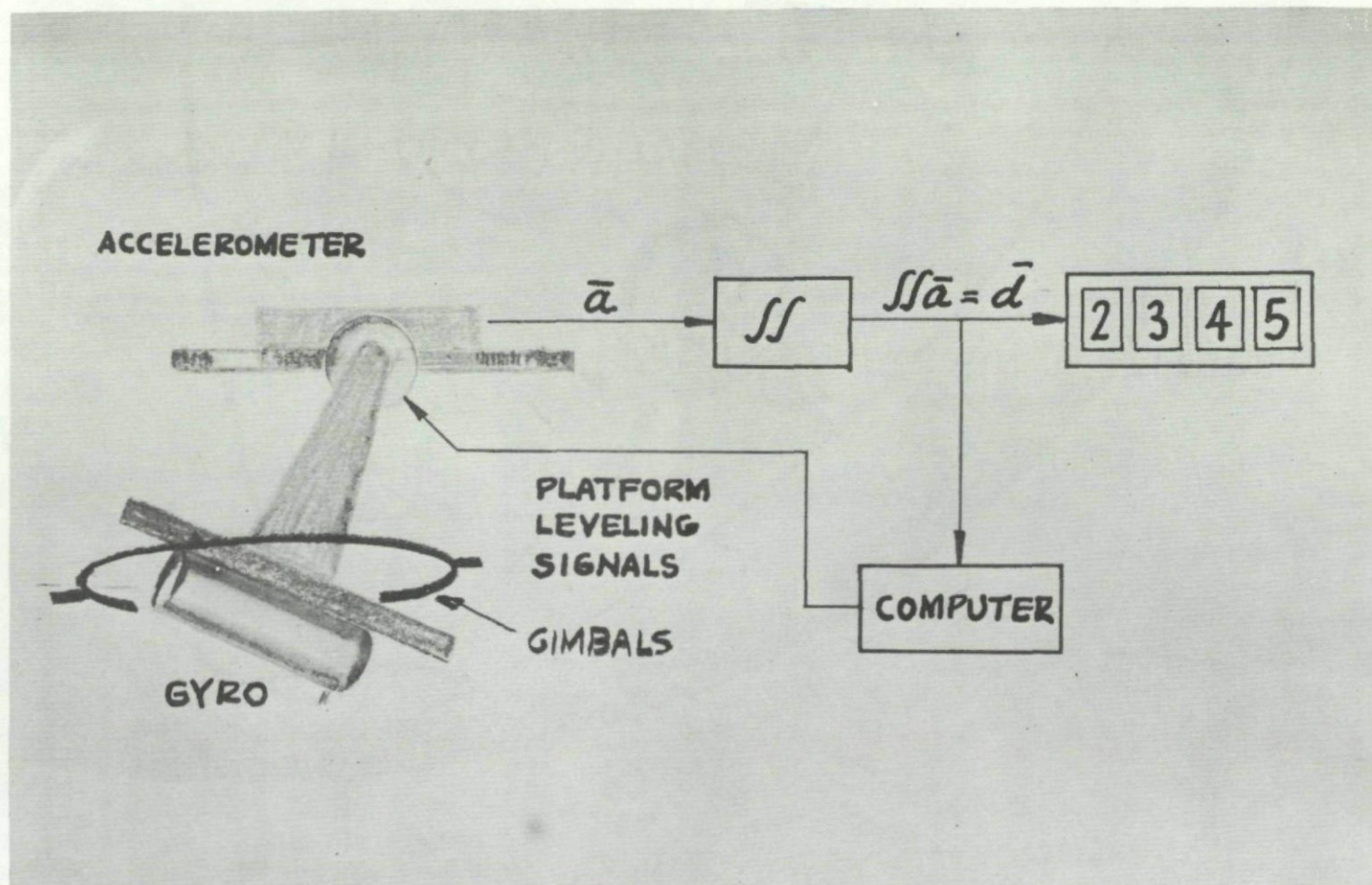


Fig. 8. System block diagram (single axis).

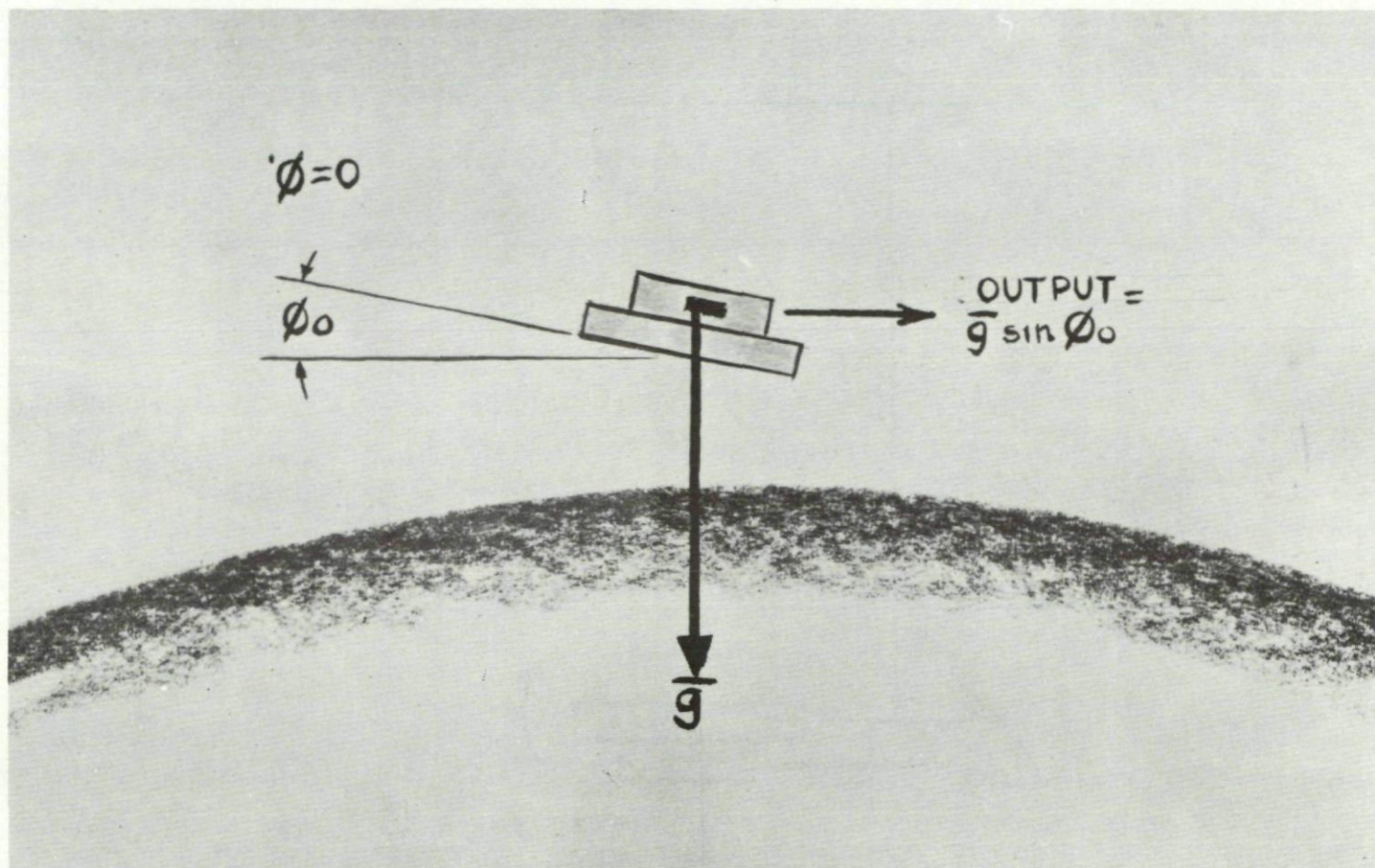


Fig. 9. Error due to initial platform tilt ($t = 0$).

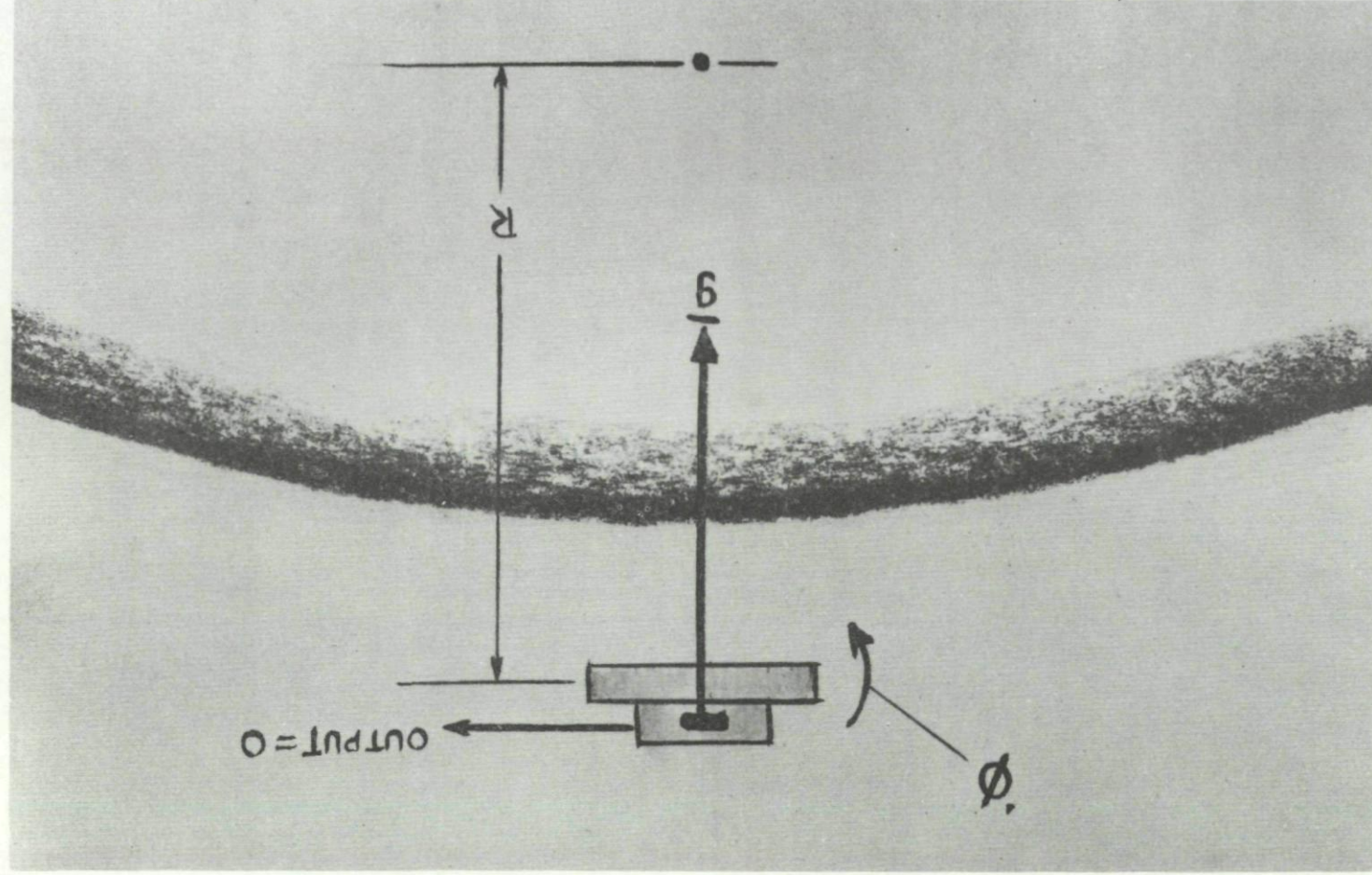


Fig. 10. Error due to initial platform tilt ($t = 21$ min).

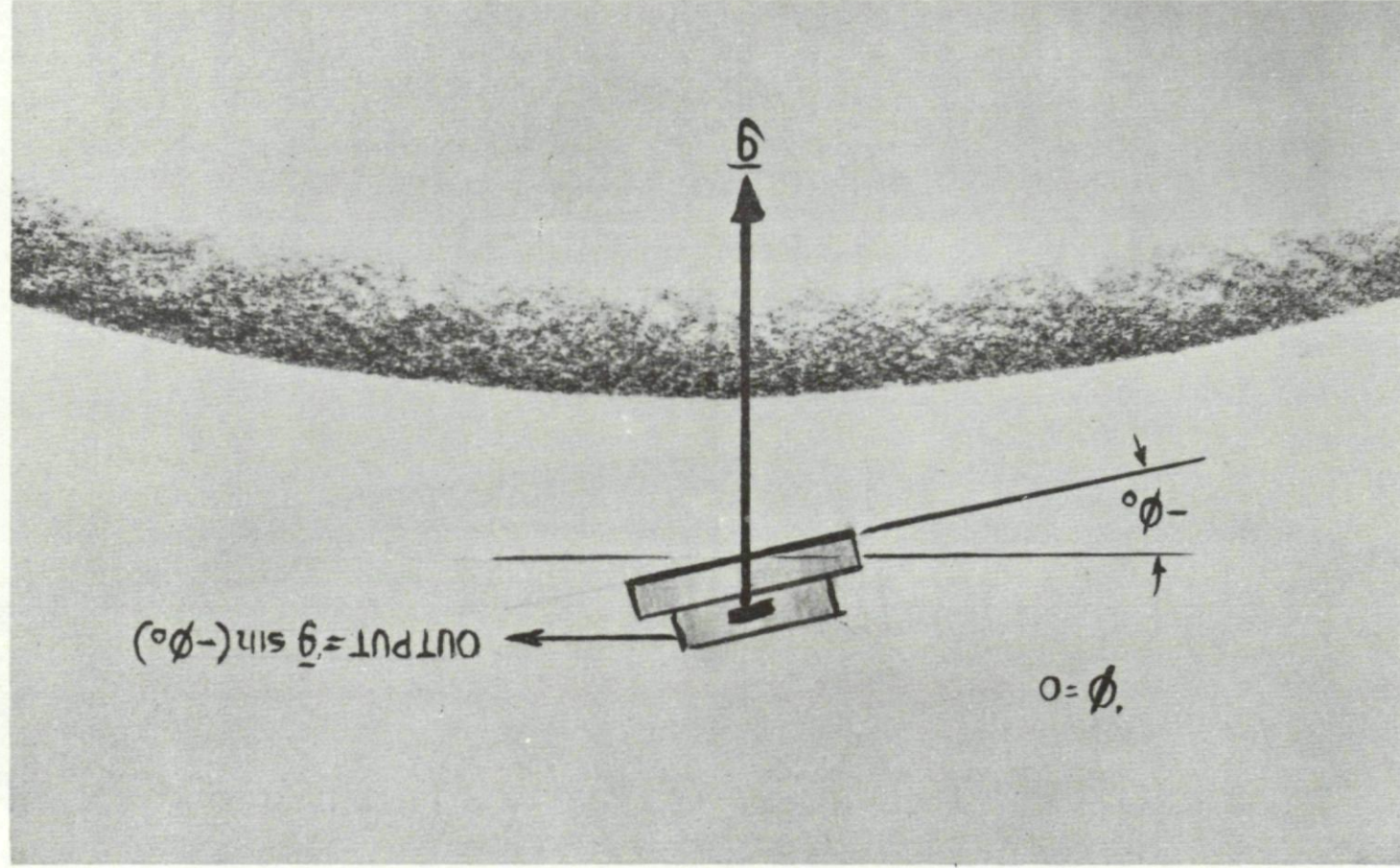


Fig. 11. Error due to initial platform tilt ($t = 42$ min).

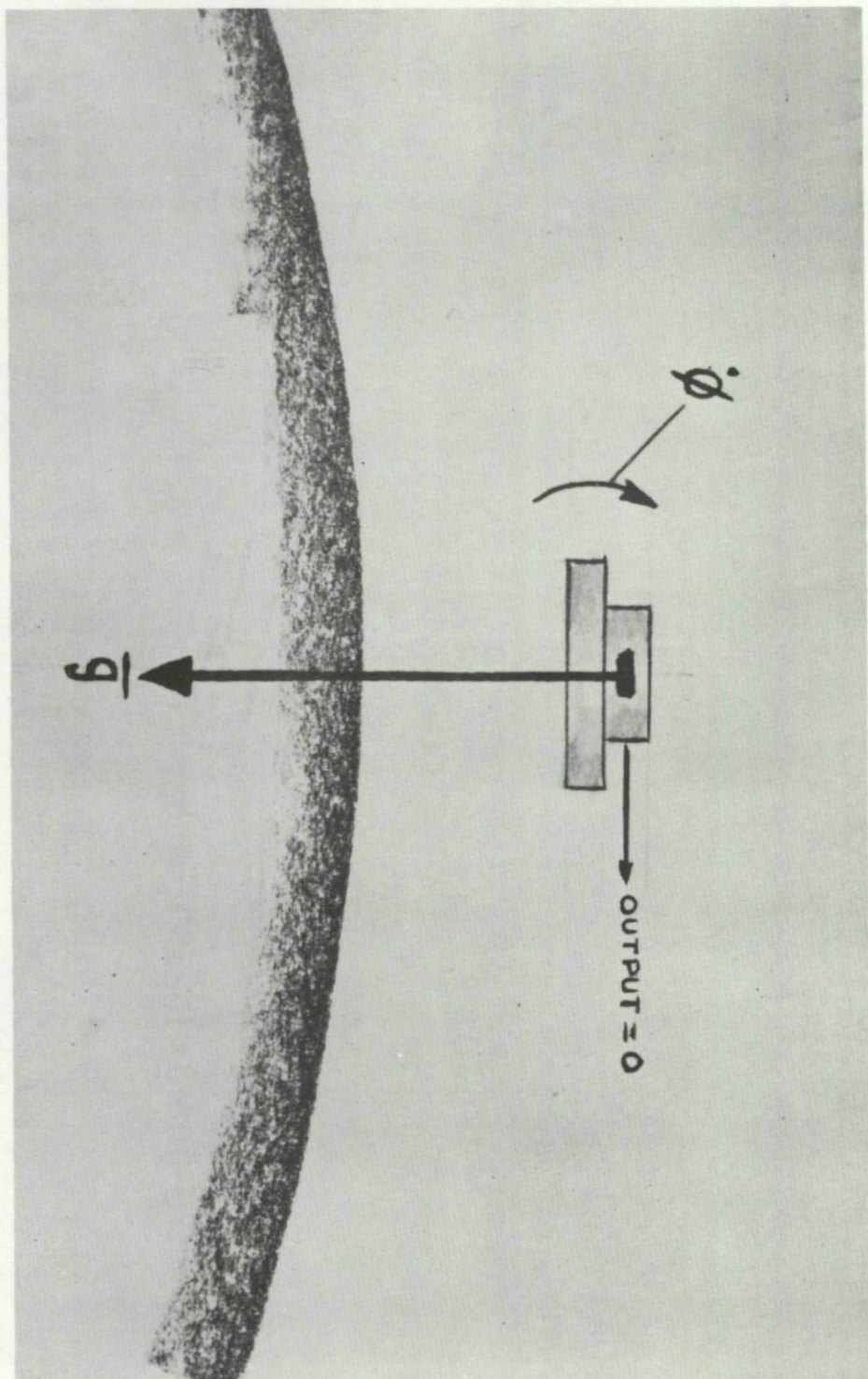


Fig. 12. Error due to initial platform tilt ($t = 63$ min).

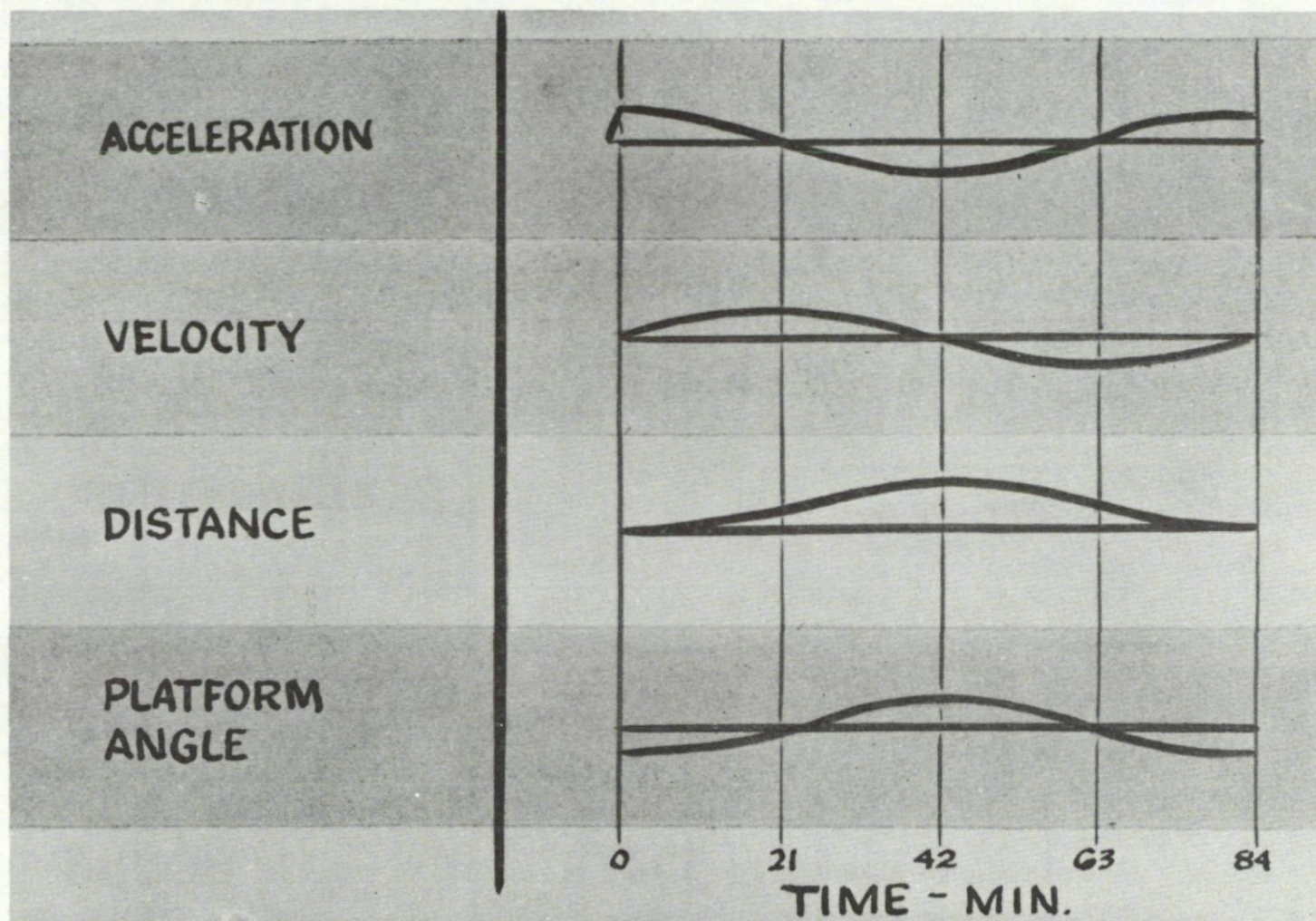


Fig. 13. Effect of initial platform tilt.

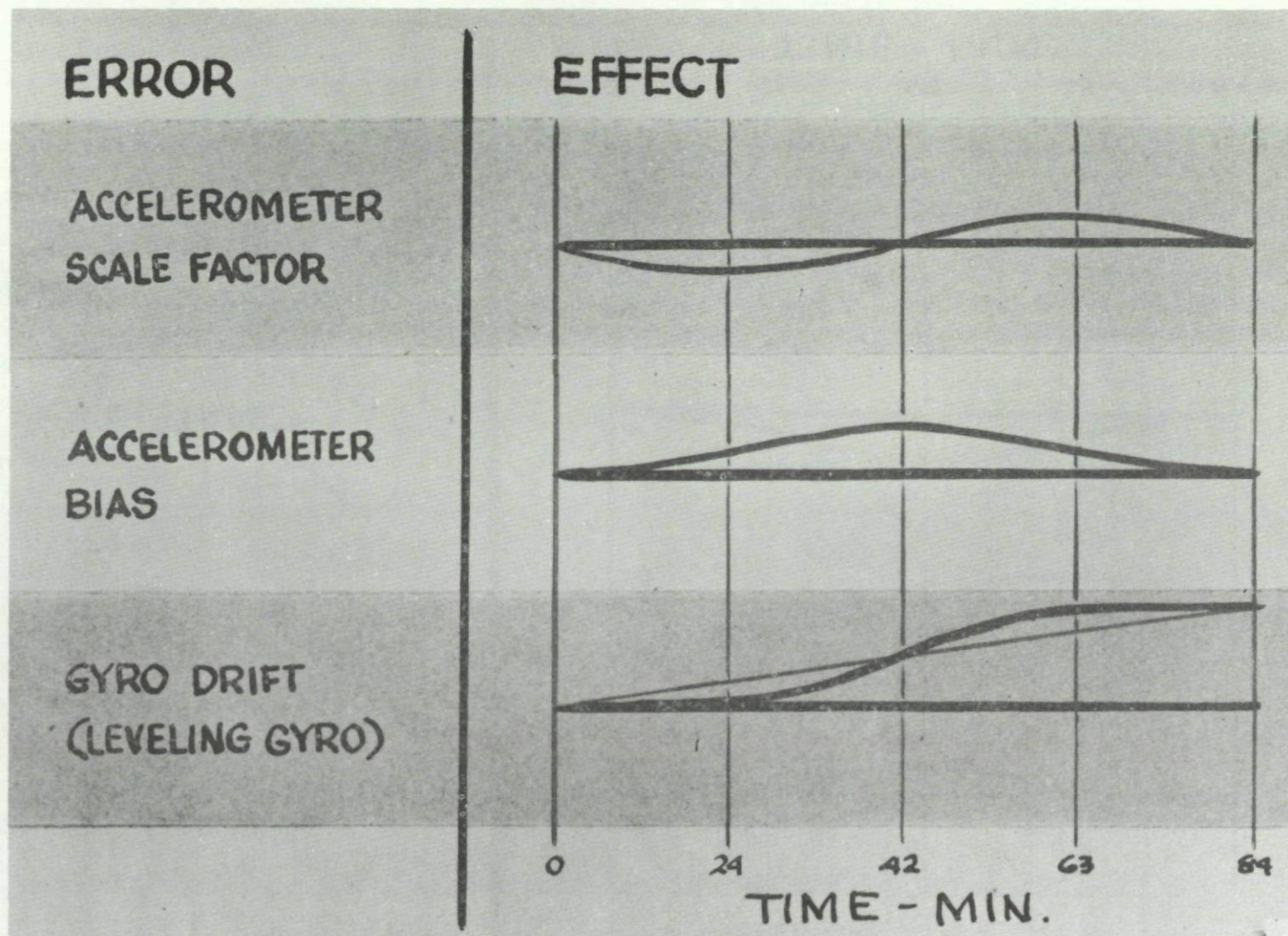


Fig. 14. Effects of various errors.

AIDING THE INERTIAL NAVIGATION SYSTEM William F. Ballhaus and Frederick Stevens, Jr.*

SUMMARY

Among the several sources of error in an inertial navigation system are gyro drift, component tolerances, mechanization approximations, deflections of the vertical, initial system misalignments, and survey inaccuracy of the launch site. While many of these errors will usually be small, others will be appreciable. The cumulative effect of errors in a long-range guidance system is such that aiding techniques will be required in order to achieve the terminal accuracies generally required of such systems. There are several aiding techniques available. Among these are the use of radar or visual checkpoints, the tracking of a radar beacon, star tracking, and computational damping through the use of an external velocity measurement. The use of two or more of these possibilities simultaneously is found to be of great value.

SOMMAIRE

Parmi les multiples sources d'erreurs dans un système de navigation par inertie se trouvent, la dérive du gyroscope, les tolérances des organes, les approximations mécaniques, les déflexions du vertical, les alignements incorrects initiaux du système et l'erreur de détermination de la position du lancement. Tandis que beaucoup de ces erreurs sont généralement petites, d'autres sont appréciables. L'effet cumulatif des erreurs dans un système de gouverne à longue distance est tel que des techniques d'aides seront nécessaires en vue d'obtenir l'exactitude finale généralement requise dans de tels systèmes. Plusieurs techniques d'aide sont disponibles. Parmi celles-ci se trouvent l'utilisation du radar ou de points d'orientation visibles, la localisation d'un feu de radar, la trace des étoiles et par amortissement calculé, basé sur l'utilisation de mesures de vitesse extérieure. L'utilisation simultanée de deux ou de plusieurs de ces possibilités se trouvent être de grande valeur.

1. INTRODUCTION

If Christopher Columbus had been wrong in the latter part of the fifteenth century, and the earth really had been flat instead of round, we would not have today's solutions to problems in geonavigation. Since Columbus was correct, and the earth is not flat, many of our problems of navigation on the earth are simplified.

In Columbus's day there was no sextant, there was no navigation almanac, and there was no speed measuring device. That he finally reached North America is a tribute to his genius, his burning desire, and his practical dead reckoning which was, with him, almost a sixth sense.

Today, with our capability to produce precision mechanism, with our superspeed electronic computers, and with our knowledge

*Northrop Aircraft, Inc., Hawthorne, California.

of the earth's geometry, it is possible to navigate precisely over long distances without reference to any landmarks, to any magnetic fields, or to any radiation fields. Inertial navigation makes this possible.

Let us consider some of the properties of inertial systems. One distinguishing property of such systems is their method of sensing present position. Here the spheroid earth permits the establishment of a one-to-one correspondence of each point on the surface of the earth with the direction of a force vector at that point. For illustration, the force vector could be the gravitational vector at that point. Fig. 1 shows how each gravitational vector is uniquely associated with a particular position on the surface of the earth. Conversely, every point on the earth's surface may be associated with the angular or directional coordinates of the vector.

It will be recognized that in precisely this manner, through the mechanism of the spherical stretch map, points on the earth's surface are identified by the classical navigator. It should be noted that these geodetic coordinates of latitude and longitude are independent of linear measurements conducted on the surface of the earth and, further, that the coordinates, (i.e., latitude and longitude) of any point may be determined solely by an angular measurement, viz: the direction of the gravitational vector observed at that point.

2. THE INERTIAL SYSTEM

The heart of today's inertial system lies in its accelerometers, supported by a stabilized platform whose space orientation is slaved to that of its reference gyroscopes. The stable platform is isolated from craft angular motions by mounting it on gimbals, involving a varying number of rotational axes, each of which represents a degree of freedom.

Accelerations are sensed and doubly integrated to derive positional information, with subsequent feedback to the inertial reference in order to isolate the effect of gravity from linear accelerations.

A diagram of the basic building blocks of an inertial system is shown in Fig. 2. In the ordinary and more elementary design, gyroscopes serve alone as the stabilization reference. Initially aligned to some known reference, they are continuously rotated in inertial space in accord with the angular velocity of the adopted coordinate system.

Any coordinate system error would introduce corresponding inaccuracy in position indication. A brief discussion of specific errors which occur in real systems and their effects upon system accuracy is necessary to orient our frame of reference.

In any real inertial system, imperfections in gyros, accelerometers, and computer response can result in measurable errors being introduced into the system. Errors resulting from such imperfections are resolved as pitch errors, roll errors, or heading errors.

A pitch or roll misalignment of one minute of arc corresponds to a direct positional error of one nautical mile on the earth's surface. A heading misalignment produces a consistently increasing positional error which increases with the distance traveled from the point where the error occurred. A heading misalignment of one minute of arc corresponds to an error rate of 0.66 miles per thousand miles traveled.

In addition to either pitch, roll, or heading error, there is a further characteristic of inertial systems which must be clearly understood.

In the simple perfect undamped inertial system, if no error is introduced, the system will always yield perfect results. However, if a specific position or velocity error is introduced into the perfect system, this error will be perpetuated in the form of an oscillating error in velocity or position as illustrated in the system in Fig. 3.

If initial errors in position e_p and velocity e_v exist, the system response is described by the Laplace transform

$$E_p(s) = \frac{e_p(0)s + e_v(0)}{s^2 + \omega_e^2} \quad (1)$$

which has the inverse transform

$$e_p(t) = e_p(0) \cos \omega_e t + \frac{e_v(0)}{\omega_e} \sin \omega_e t. \quad (2)$$

This oscillation is similar to the oscillation of a pendulum whose length equals the earth's radius; its characteristic natural period, determined by gravity and the effective pendulum length, is 84.4 minutes.

It can be readily seen too, in Fig. 4, how gyroscopes and reference misalignments enter system dynamics as major offenders among the many possible sources of system error.

For example, gyroscopes whose drift rate is as low as .01 degree per hour can permit an error growth at a rate of almost one mile per hour, quite apart from any misalignments which were permitted to exist prior to initial system operation. Even this precision is insufficient to meet accuracy requirements in some long-range operations regardless of perfection in the remainder of the system.

Insofar as initial computer settings and initial alignment of the stable reference are concerned, little difficulty is encountered in ground-based systems, where extensive surveys and fixed references may be employed to yield almost any practical desired accuracy. However, such facility is not present in systems operated from moving bases such as ships at sea or airborne aircraft, and accuracy specifications may in this event prove too demanding to be met by conventional concepts of stable platform mechanization and system alignment. In addition, some ground-based systems which must be placed into action on short notice do not permit sufficient time for satisfactory preflight operations.

Accelerometers of demonstrated range, sensitivity, and linearity sufficient to justify the use of a simple undamped system have been unavailable in time to be used for some navigation systems. The required sensing range presents a particularly acute problem inasmuch as it affects achievement of the other desired characteristics. It is clear that systems of the type under discussion may experience large accelerations both during takeoff and in flight maneuvers. In naval installations, large transients may also be caused by the ship's natural pitch and roll.

Integrator response presents still another problem in the computational aspect of inertial navigation. Here, any inaccuracy is immediately reflected in an equivalent error of integration, which continues to be propagated within the undamped system in its characteristic 84-minute oscillation. This can be particularly offensive in a total coordinate computation, in which integrator range must be spread over that of the necessary flight range. Most approaches are confronted with difficulty which is encountered in accelerometer design - that of achieving both wide range and high sensitivity in an analog device.

It is evident that the unaided inertial system may easily prove unable to perform satisfactorily within operational restrictions and ranges associated with a number of applications in which it could otherwise serve to real advantage. To the designer confronted with this situation, who desires to retain the advantages of inertial navigation, several alternatives are at once apparent. The first is to proceed with a simple inertial system in the hopeful belief that eventual development of acceptable accurate production components will be possible within the allotted time span. A second, and more conservative course involves augmenting the basic system in such a manner as to eliminate its significant limitations as a result of component error and at the same time to escape the risks involved in expecting results too soon in difficult and uncertain development areas.

3. AIDED INERTIAL SYSTEM

Of the several possible methods of aiding an inertial system, incorporation of checkpoint information usually, and understandably, is given first priority. This is especially true in piloted aircraft applications, in which the crew can be of assistance, because of its obvious direct relationship to the prime navigational output. Search radar and visual systems have been employed in this respect, the former for its advantage of area coverage, the latter for relative definition and precision. Of course, some reasonable distribution of checkpoints must exist in or near the area where accuracy is of interest if useful information is to be derived.

It is possible to automatize the checkpoint operation. The fashion in which checkpoint information usually is introduced to the basic inertial computer to effect the necessary correction is quite elementary, and its effect is illustrated in Fig. 5. As shown,

residual velocity error, unaffected by correction of observed position error, is left to propagate a new, but not larger, error. Thus, a series of checkpoint observations at suitable intervals can lead at each step to a further reduction of both position and velocity uncertainty, as shown in Fig. 6.

These corrections, when applied specifically to the computer, permit any reference misalignment to continue as a source of position error. Other and somewhat more complex approaches to checkpoint correction which affect reference as well as computer alignment are feasible.

Airborne radar, tracking a beacon whose position is known accurately, can provide a rapid reduction of errors in both position and velocity; this essentially continuous form of checkpoint incorporation can be performed equally well with interchanged location of tracker and beacon, if a communication link is added to the system. Both velocity and position corrections are effected, with time constants chosen on the basis of a compromise between the desire for rapid decay of the error and the necessity to avoid unpleasant effects of tracking noise. When the beacon can be tracked at short range an appreciable increase in precision can be achieved.

A final method of checkpoint incorporation involves the tracking of stars, a procedure familiar to any air force or navy navigator. If the direction of the vertical is assumed to be known, as well as an accurate indication of time, the "shooting" of stars can establish local position. This procedure lends itself to automatic methods, with the reference platform serving to indicate the vertical.

Perhaps the most important consideration in connection with reliance upon checkpoints is the distance to be traveled from the last checkpoint to the target, since the inertial

reference (and computer) can easily be drifting during the time required to traverse this final leg of the flight. The problem which this presents is, of course, a function of craft speed as well as of the distance involved.

When operational conditions do not permit accurate alignment prior to takeoff, and when available gyroscopes can be expected to allow excessive drift of the stable platform, additional equipment to correct the stable platform is necessary.

To meet this problem, the star-tracking equipment mentioned previously can be used with a different objective. One or more automatic star trackers can be used in connection with the gyro-stabilized platform to establish the platform space orientation.

In operation, this utilization of star tracking provides signals to the gyroscopes which serve to correct their angular velocity as well as their angular position in inertial space. The stabilized platform continues to operate just as it does in the basic inertial system, with the exception that its performance is enhanced; gyroscope drift is eliminated, in effect, as are misalignments due to severe launch shocks and to poor initial conditions such as inexact knowledge of the launch site coordinates. During any absence of star-tracking information, the gyro-stabilized platform performs just as it would have in an unmonitored system. Star tracking can be initiated anew after periods of cloud cover, within limits of the design parameters of the system such as telescope field size and expected drift rates of the unmonitored platform.

Auxiliary telescope pointing, stellar detection, and stellar data resolution equipment completes the star-tracking system. A number of approaches can be followed in its

design, whose basis may range from a general purpose computer to a playback of precomputed telescope orientation data. The latter approach is favored for its reliability, the former for its flexibility.

The utilization of an independent source of velocity information for improving inertial navigation system performance through the damping of dynamic errors is evident. Let us consider three modes of utilizing velocity information.

We may process the independent velocity information in such a manner that it may be inserted into the computer as (1) a direct correction to the computed linear velocity, (2) a derived correction to the measure of linear acceleration, or (3) a derived correction to the computed position.

The mathematical equivalence of these three approaches has been established for linear filter design. Because of this equivalence and the directness of the mode wherein the auxiliary velocity information is used to correct the computed velocity, this mode will be used as a basis of further description. A general functional diagram is given in Fig. 7. The conditions which determine the finalized filter design for any particular system are:

- a. The simplicity of mechanization desired.
- b. The character of the noise associated with the measurement of velocity.
- c. The desired performance.

Generally the desired performance is damping of the 84-minute oscillation associated with the initial conditions, and attenuation of the effects of noise in the accelerometer-computer loop. The complexity of the velocity filter is usually restricted to first-order operations and not more than two

parameters. The selection of the optimum magnitude of these parameters is uniquely determined by the desired rapidity of damping and the nature of the velocity measurement. The current methods of velocity measurement are by Doppler shift in radar echoes and by true airspeed. In the latter approach, wind variations must be lumped with other velocity noise for purposes of determining the filter parameters.

Velocity damping removes only errors in position and velocity associated with the 84.4-minute oscillations. This is shown by deriving the Laplace transform of the positional error for the system in Fig. 7. This can be shown to be

$$E_p(s) = \frac{e_p(0)s + e_v(0) + Ks^2 N(s)}{s^2 + Ks \omega_e^2 + \omega_e^2} \quad (3)$$

where K is a filter parameter and $N(s)$ is the velocity noise function.

Inverting this transform gives the time response

$$e_p(t) = e^{-\alpha t} \left\{ e_p(0) \left[\cos \beta t - \frac{\alpha}{\beta} \sin \beta t \right] + \frac{e_v(0)}{\beta} \sin \beta t \right\} + e_N \quad (4)$$

where e_N is the noise-forced error, α is $K \omega_e^2 / 2$ and β is $\sqrt{\omega_e^2 - \alpha^2}$.

Fig. 8 further illustrates the fact that velocity damping does not correct errors due to platform misalignment.

To this point, the discussion has been concerned with the effects of individual techniques for aiding the inertial system. It has been shown that some of these schemes correct the computer, while others secure proper platform orientation; in some cases, either the computer or the platform may be aided, depending upon the manner in which the auxiliary information is introduced into the basic system.

This is accomplished through a combination of two or more of the possibilities already enumerated. For example, the platform orientation could be corrected or improved by radar checkpoint with the computer employing velocity damping. The combination of star tracking for more exact platform orientation with radar checkpoint corrections for the computer is also possible. Highly satisfactory results could be obtained combining star tracking and velocity damping. The effect of the tracking in this case would be to insure that the computer is damped to the proper reference, so that the computations would not be merely nonoscillating, but actually correct.

4. CONCLUSION

It is important to note that inertial navigation systems are no mere mathematical concepts. They are living, breathing, working elements. They come in all sizes and shapes, with accuracies to match. Every type of aiding scheme discussed here is either a part of an existing system or has at least been flight-demonstrated. Star tracking is as feasible in daylight and twilight hours as in the dark of night; checkpoints can be used for computer correction; velocity damping works as theory indicates it should. Some systems use single aids, some apply combinations, and others get along with no aids at all. The state of the art is such that an inertial navigation system either has been built or could be designed to meet any reasonable set of requirements if size is not restricted. Development effort is now directed to the reduction of size and weight of inertial systems.

How simple Christopher Columbus's job of discovery of North America and return to Spain would have been if he could have installed an automatic inertial guidance system in his Nina, Pinta, or Santa Maria!

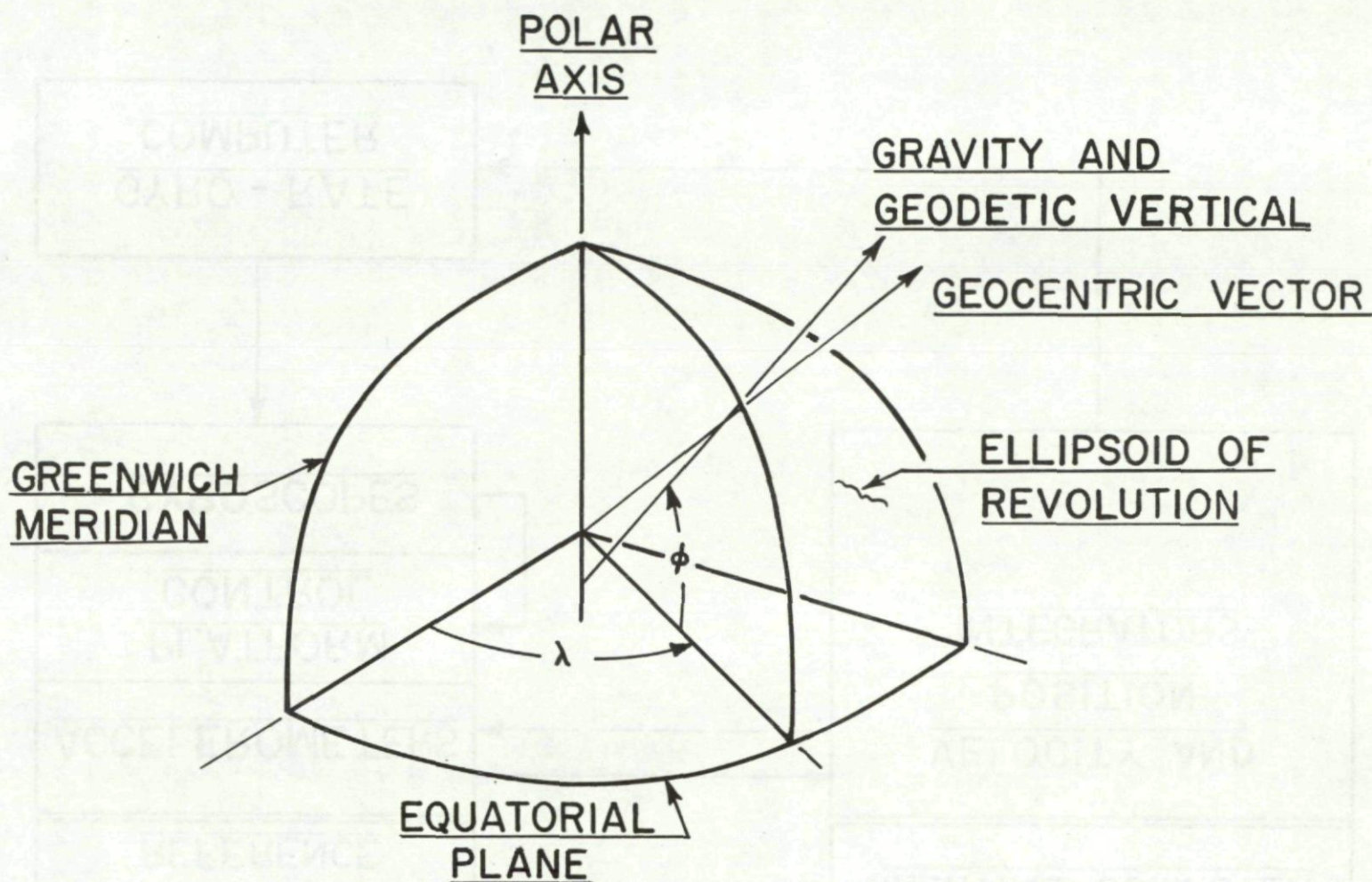


Fig. 1. Relationship of gravitational vector and points on surface of the earth.

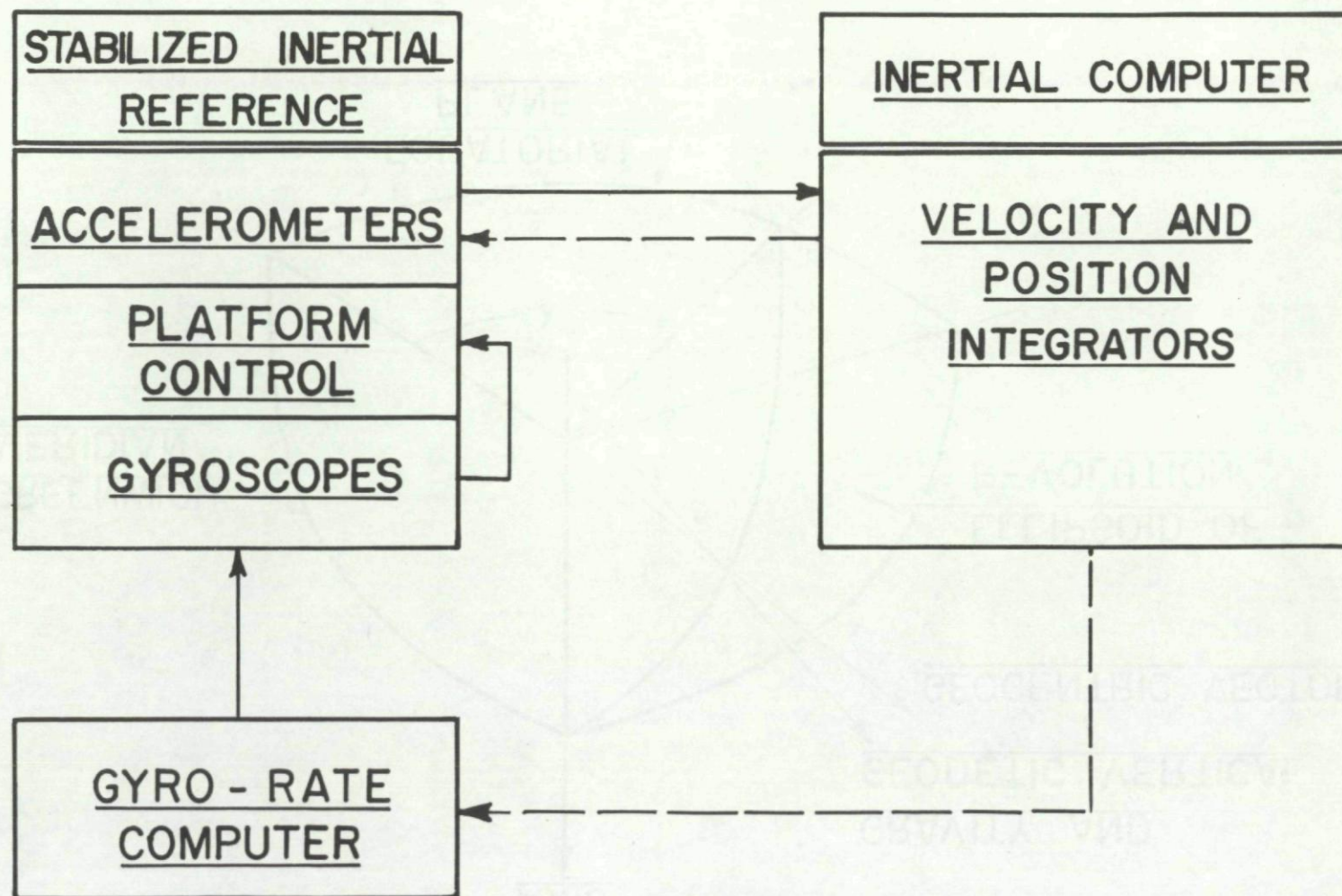
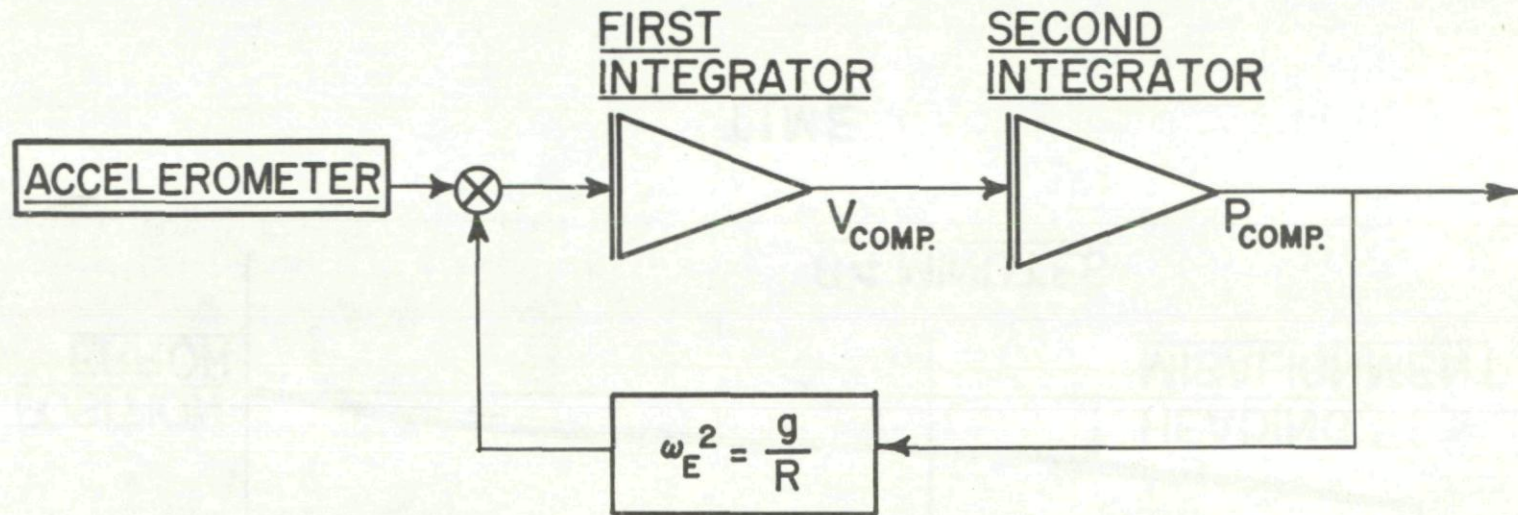


Fig. 2. Basic inertial system.



95

SYSTEM RESPONSE TO INITIAL CONDITION ERRORS

TRANSFER FUNCTION

$$\epsilon_p(s) = \frac{\epsilon_p(0)s + \epsilon_v(0)}{s^2 + \omega_E^2}$$

$\epsilon_p(0)$ = INITIAL POSITION ERROR

$\epsilon_v(0)$ = INITIAL VELOCITY ERROR

$$\epsilon_p(t) = \epsilon_p(0) \cos \omega_E t + \frac{\epsilon_v(0)}{\omega_E} \sin \omega_E t$$

Fig. 3. Undamped inertial computer.

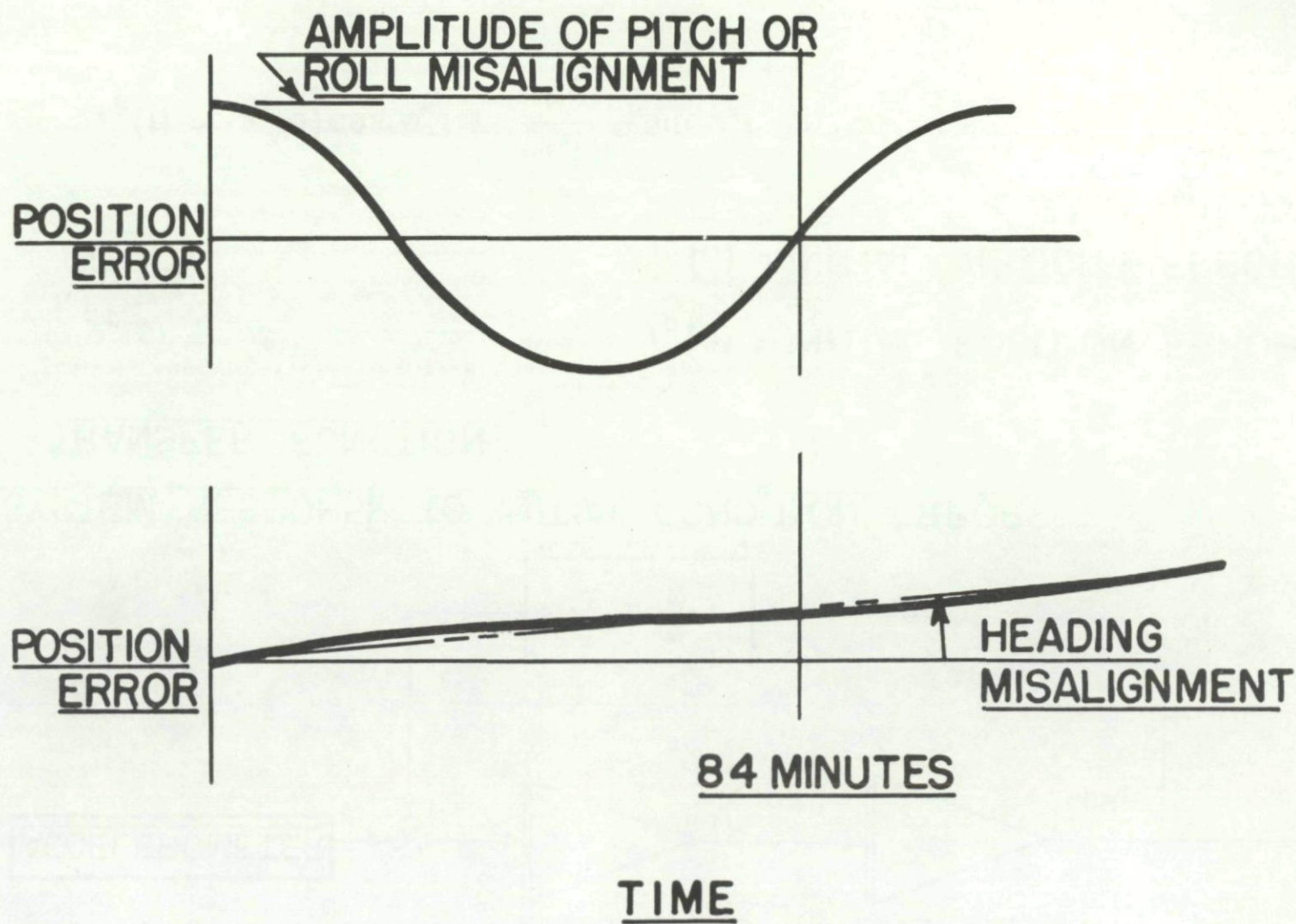


Fig. 4. Undamped system response to reference misalignment.

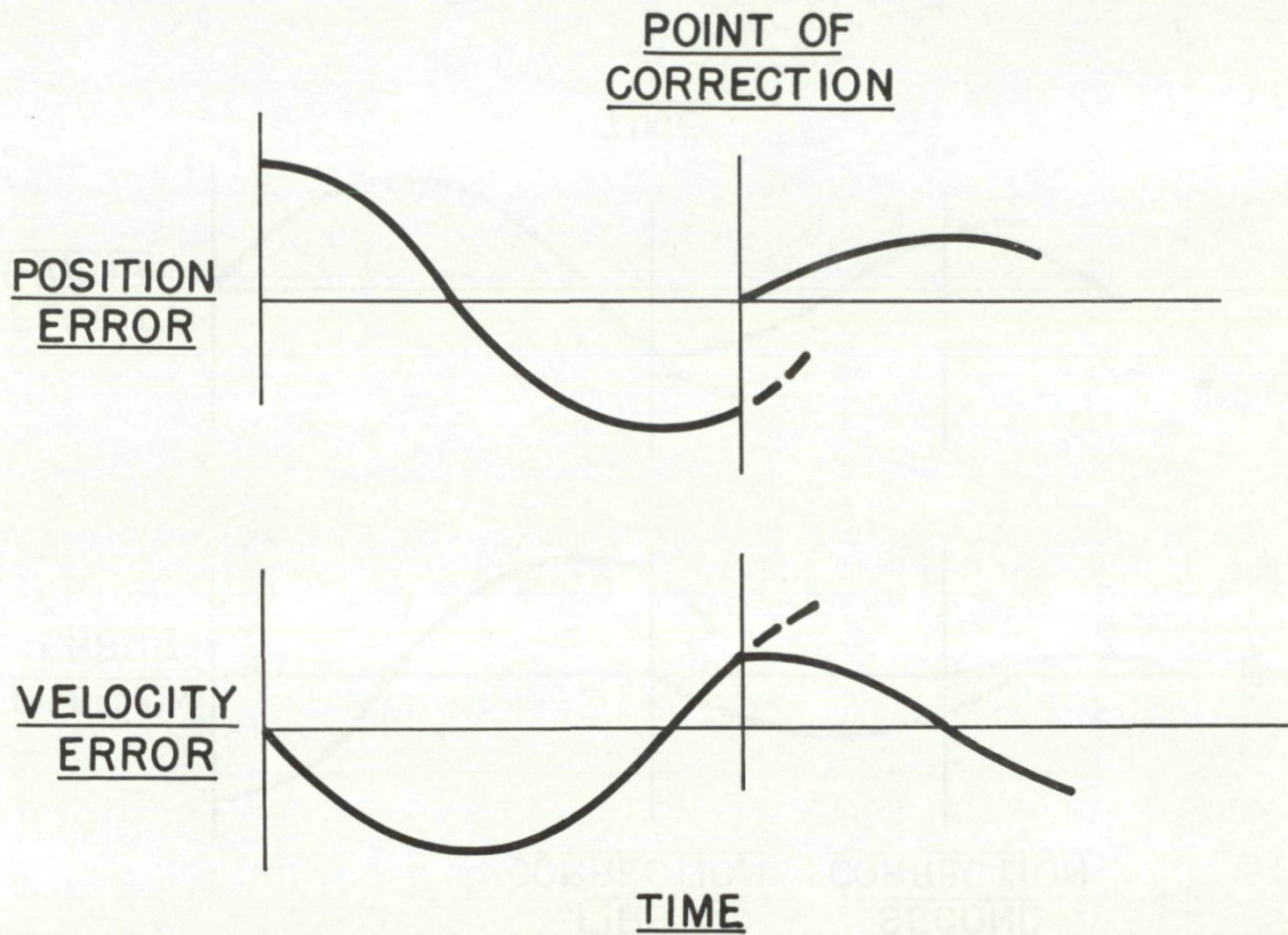


Fig. 5. Effect of single checkpoint incorporation.

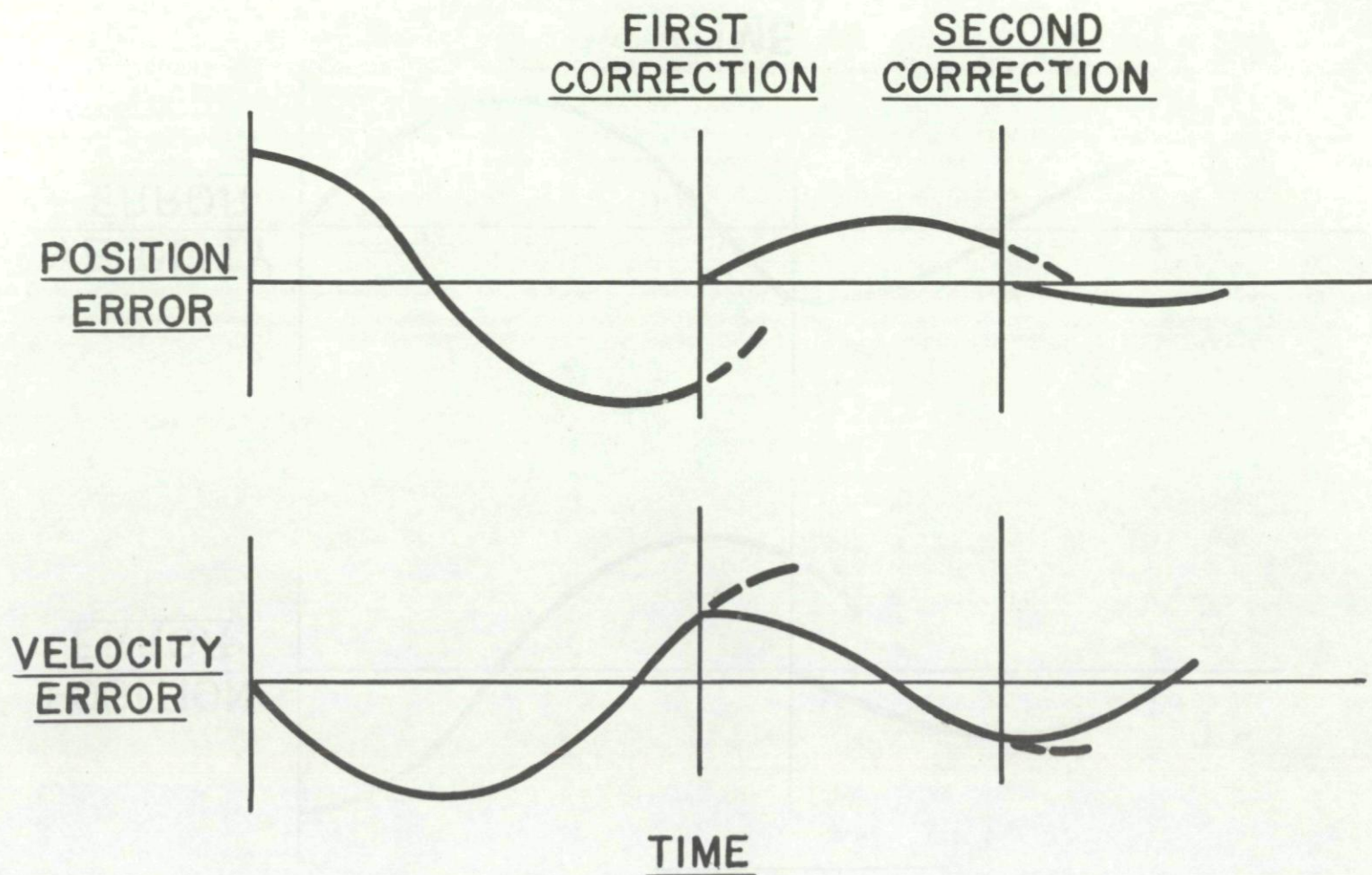


Fig. 6. Effect of repeated checkpoint incorporation.

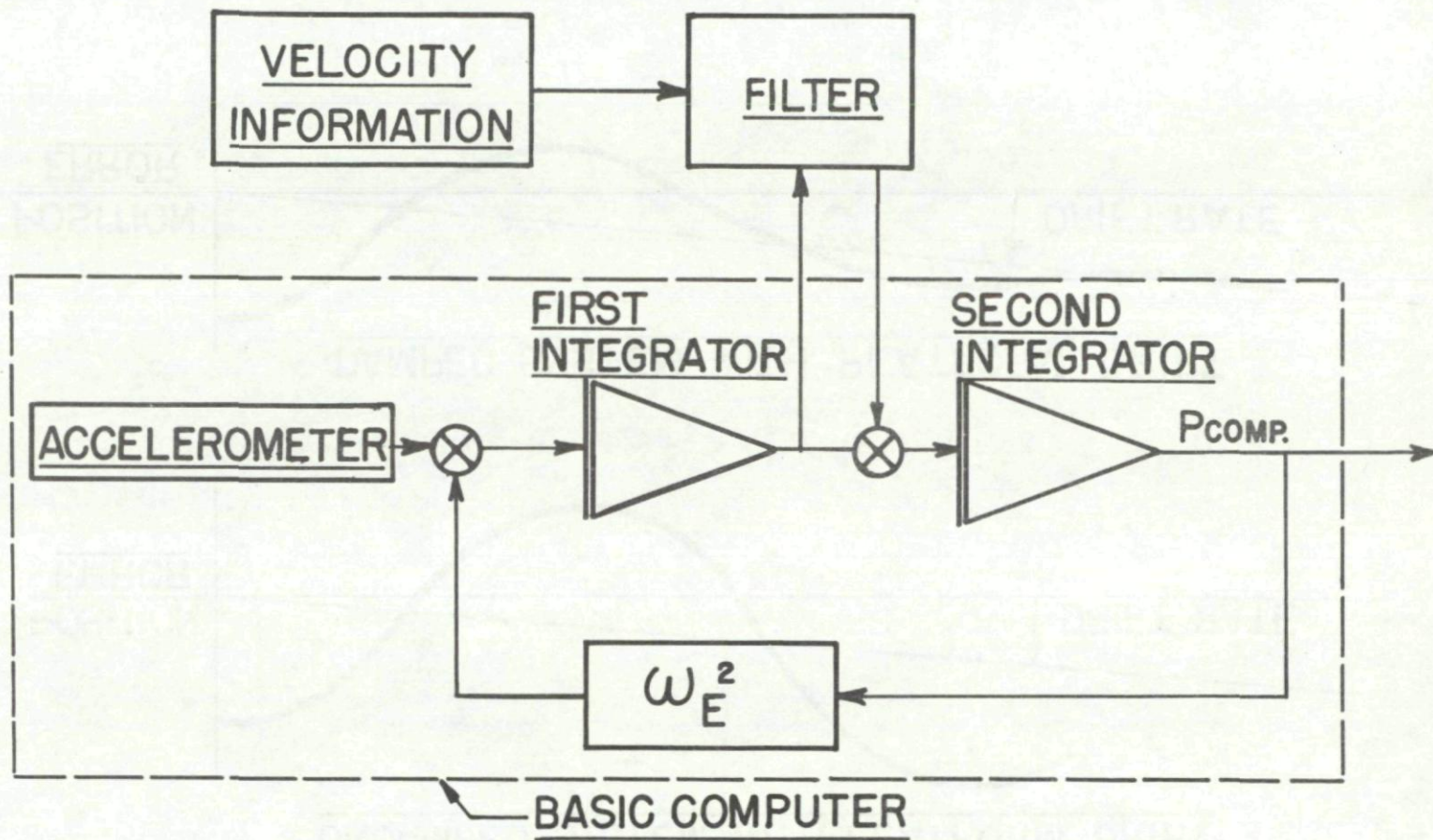
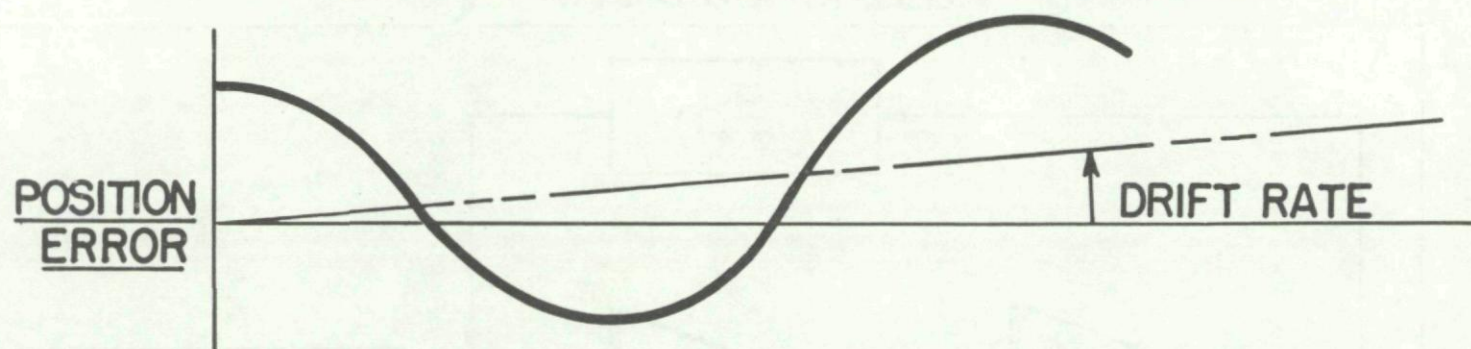


Fig. 7. Velocity-damped inertial system.

UNDAMPED SYSTEM WITH PLATFORM DRIFT



DAMPED SYSTEM WITH PLATFORM DRIFT

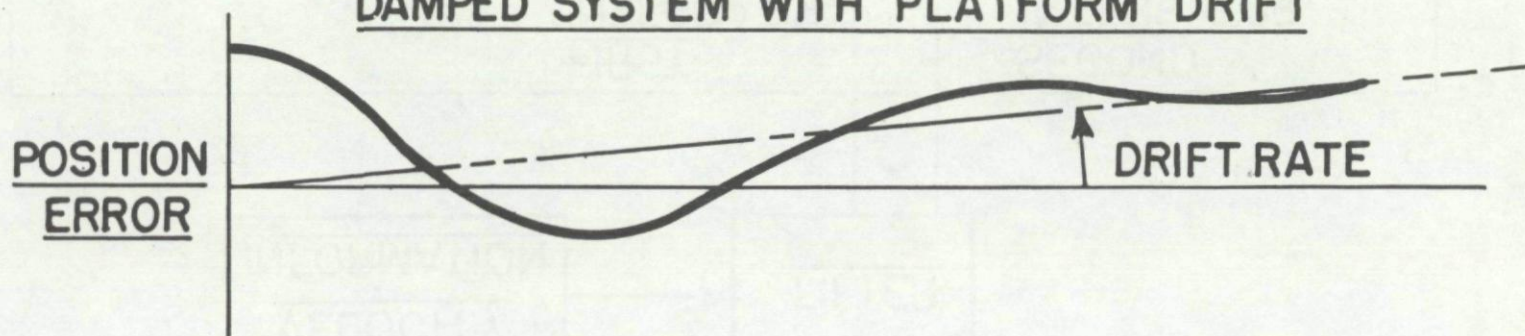


Fig. 8. Response of damped computer to initial conditions.

LINEAR HOMING NAVIGATION

Robert K. Roney*

SUMMARY

This paper deals with the analysis of proportional navigation. The equations of navigation are developed and the effect of system dynamics and smoothing time on the resulting trajectories and miss distances is briefly discussed.

SOMMAIRE

Cette note est une analyse de la navigation proportionnelle. Les équations de navigation sont écrites et les effets des éléments dynamiques du système et du temps d'atténuation sur les trajectoires et les distances manquées sont brièvement discutés.

1. INTRODUCTION

Linear homing navigation is a special case of line-of-sight navigation. By line-of-sight navigation is meant navigation using information on the bearing only of the intended target, i.e., without use of the range. A familiar example of such navigation is the use of a radio direction finder in an airplane to home on a radio beacon. This discussion will be limited to guidance systems in which the rate of rotation of the velocity vector of the guided vehicle, hereinafter called the missile, is made linearly proportional to the observed rate of rotation of the line-of-sight to the target, in such a way as to reduce the line-of-sight rotation. When the rotation of the line-of-sight is zero, the missile is on a linearly predicted collision course with its target. Such guidance schemes are generally called proportional navigation.

Consider the coordinate system of Fig. 1. Perfect proportional navigation is described by the equation

$$\dot{\gamma} = \lambda \dot{\sigma} \quad (1)$$

where γ represents the angle of direction of the missile, σ represents the angle of the line-of-sight from missile to target, both measured with respect to some fixed reference; the dot indicates differentiation with respect to time, and λ is a constant called the navigation gain.

It will be recognized that $\dot{\gamma}$ represents a lateral acceleration a_m , of the missile and that $\dot{\sigma}$ arises from a component of the relative velocity normal to the line-of-sight, i.e., a velocity deviation from a collision course, thus

$$a_m = V \dot{\gamma} \quad (2)$$

where V is the missile speed and

$$\dot{\sigma} = \frac{d}{dt} \left(\frac{\bar{V}_c - \bar{V}}{R} \right) \quad (3)$$

where \bar{V} is a vector representing missile velocity and \bar{V}_c the velocity required for a collision course.

*Hughes Aircraft Co., Culver City, California.

With this interpretation firmly in mind, the conception is easily extended to three dimensions wherein, for the most efficient navigation, the component of the missile lateral acceleration normal to the line-of-sight is made in the negative direction of the velocity error. (See Ref. 1.) In other words, the missile acceleration is in the plane of rotation of the line-of-sight. This requires that the navigation gain, as defined by Eq. (1), in the two principal planes be the ratio of the cosine of the lead angles.

An elegant notation of the general three-dimensional law is the vector equation

$$\dot{\bar{V}} = \lambda \bar{\omega}_\sigma \times \bar{V} \quad (4)$$

in which $\bar{\omega}_\sigma$ represents the rotational velocity vector of the line-of-sight.

With this generalization, it can be shown that the resultant trajectory equations can, with appropriate linearization, be separated into two identical independent scalar equations. To simplify the discussion, henceforth we may consider only the two-dimensional case, or one of the independent equations.

Returning to the basic equation, it should be pointed out that any actual mechanization must result in modification of the equation to account for the dynamics of the situation. Generally speaking, any measure of the line-of-sight direction σ made within the missile will include random errors, or noise. Under most practical conditions, elimination or reduction of this noise will be required by filtering, or smoothing, with a resultant delay in signal determined by the smoothing or averaging time. In addition there will be unavoidable dynamic delay in the basic control equipment with which the required accelerations are developed.

Including these dynamic terms and noise, the control equation becomes

$$Z(p) \dot{\gamma} = \lambda \dot{\sigma} + \lambda \dot{\eta} \quad (5)$$

where $Z(p)$ is a polynomial (or ratio of polynomials) in the differential operator $p = d/dt$, with static value unity, characterizing the system dynamics; $\dot{\eta}$ represents errors in the measurement of $\dot{\sigma}$.

2. BASIC TRAJECTORY SOLUTIONS

Perhaps the best picture of the general trajectory relations arising from proportional navigation is revealed by direct solution of the proportional navigation equation without system dynamics, i.e., $Z(p) = 1$. This can be done by noting the relation between γ and σ inherent in the missile-target kinematic system (Fig. 1), from which one observes that

$$\begin{aligned} R\dot{\sigma} &= U \sin(\phi - \sigma) - V \sin(\gamma - \sigma) \\ \dot{R} &= U \cos(\phi - \sigma) - V \cos(\gamma - \sigma) \end{aligned} \quad (6)$$

where R is the distance between missile and target. If V and U are constant, it follows from Eq. (6),

$$\ddot{\sigma} = -2 \frac{\sigma \dot{R}}{R} + \frac{\dot{R}_m}{R} \dot{\gamma} + \frac{a_{tp}}{R} \quad (7)$$

where a_{tp} is the target acceleration normal to the line-of-sight.

When combined with the control equation (5), a first order differential equation in $\dot{\sigma}$ is obtained which can be solved directly yielding

$$\lambda \dot{\sigma} = \dot{\gamma} - \lambda \dot{\eta} = \lambda \dot{\sigma}_0 \left(\frac{R}{R_0} \right)^{(\Lambda-2)\dagger} + \frac{\Lambda}{\Lambda-2} \left[1 - \left(\frac{R}{R_0} \right)^{(\Lambda-2)\dagger} \right] \left(\frac{a_{tp}}{V} - \dot{\gamma}_\epsilon \right) \quad (8)$$

where $\Lambda = \lambda \dot{R}_m / \dot{R}$, the effective navigation gain, and $\dot{\gamma}_\epsilon = \lambda \dot{\eta}$, the measurement error in terms of $\dot{\gamma}$.

The important feature which should be observed in this equation is the effect on the rate of error decay and required missile acceleration by the navigation parameter Λ . Note in particular that as Λ approaches 2, the acceleration required to correct initial errors does not decay, and the terminal acceleration required by evasive target maneuvers becomes unbounded. The two terms of the trajectory equation are shown in Figs. 2 and 3 for various values of Λ .

In order to minimize the total trajectory curvature, it is clearly desirable to make the constant λ as high as allowed by other considerations. On the other hand, as may be discerned from the basic control equation, the magnitude of λ is limited by the allowed accelerations resulting from noise or random inputs, and by stability considerations arising out of spurious feedback mechanisms in the missile system.

Figs. 4 and 5 illustrate the trajectories described by Eq. (8) for various Λ . For clarity, these are drawn for a stationary target. Noting that drift or unbalance in the missile control equation is equivalent to a target acceleration always normal to the line-of-sight, the trajectories for both initial error and steady maneuver can be shown without motion of the target point.

3. MISS EQUATIONS AND WEIGHTING FUNCTION

When the control dynamics, i.e., the time delays, are included, the trajectory equations cannot in general be solved explicitly. They can, of course, be plotted by simulation techniques, or integrated numerically with digital computers. Fortunately, details of the actual trajectories are seldom required if the general characteristics of the trajectories are known. The main point of interest is the closest approach, or miss distance, to the target. By application of the initial value theorem to the Laplace transform of the trajectory differential equation, it can be shown that there is a formal analytical solution of this miss distance in perfect generality for any transfer function for the control dynamics. (See Ref. 2.) This solution can be applied to any initial conditions and any disturbances such as target maneuver or noise.

The general solution, however, is rather formidable, as indicated by the formal equation for the miss:

$$m(o) = y_t(o) - \frac{1}{2\pi i} \oint g(s) Y_t(s) ds \quad (9)$$

or

$$= \mathcal{L}^{-1} \left[g(-s) Y_t^*(-s) \right]_{T=0} - y_\eta(o). \quad (10)$$

The integral will be recognized as the residue at infinity of the integrand. \mathcal{L}^{-1} is the inverse Laplace transform. $Y_t(s)$ is just the Laplace transform of the target evasion and noise driving functions, and Y_t^* is the same transform with the contribution of the discontinuity at $T = 0$ omitted; $g(s)$ is analogous to the system function in ordinary constant-coefficient linear system analysis,

and is readily calculated from the following equation if the control transfer function $Z(p)$ is known in factored form:

$$g(s) = \exp \left[\Lambda \int_{-\infty}^{-s} d\rho / \rho Z(\rho) \right]. \quad (11)$$

Asymptotic solutions of the formal miss equation are fairly readily calculated, however, for particular cases of interest. The form in which the solutions are of most interest is the so-called weighting function, i.e., the miss resulting from a specified momentary disturbance as a function of the time or range at which the disturbance is applied. From suitable linear combinations of these weighting functions, the miss arising out of any conceivable target evasive motion, measurement error (noise), or initial conditions may be determined.

While the weighting functions can be computed in the fashion discussed, it is generally far more efficient to plot them directly with a simulator using the adjoint technique on the trajectory equations. A few such weighting functions are shown in Fig. 6.

It is convenient to normalize the weighting functions in terms of some characteristic time constant of the missile system, usually the total smoothing time. Particular attention is called to the weighting functions for velocity error (launching error) and target acceleration. Note particularly that at sufficiently long flight time, for a navigation constant at least as great as three, the miss resulting from initial errors or steady target maneuvers (and hence from internal system drifts, or zero errors) vanish. Another way to state this observation is that for any specified flight time, the miss resulting from these errors vanishes for sufficiently small smoothing time. The time required to become insensitive to initial or steady disturbance is relatively insensitive to the system gain, or navigation constant.

4. OPTIMIZATION OF DESIGN PARAMETERS

There are two main parameters characterizing a linear homing system, the navigation constant and the smoothing time. Similarly there are two performance characteristics of prime interest that are affected by the design parameters: the miss distance, and the maximum acceleration experienced. We are generally interested in these characteristics as stemming from four sources: initial conditions, target evasion, noise, and system errors, e.g., drifts or zero shifts. We shall now see how these factors are related in a general way in the optimization of such a system.

The acceleration required to counter steady evasive maneuvers or system zero errors, provided the flight time is long compared with the smoothing time, varies as $\Lambda / (\Lambda - 2)$ (c.f. Eq. (8)), and thus diminishes as the navigation gain increases. On the other hand, as we can readily see from the basic control equation, the acceleration resulting from random, short period, error signals on line-of-sight measurement gives accelerations, at least at long range, proportional to the navigation gain, but reduced by the attenuation of the smoothing filters which depends on the exact nature of the filter and the spectral frequency of the random noise.

In general, the resulting acceleration will be inversely proportional to the square root of the smoothing time. Looking at the gain effects first, Fig. 7 illustrates the trade-off between acceleration requirements for countering target acceleration and noise as a function of Λ . Note that for a specified target acceleration and a specified magnitude of noise power density and smoothing time, there is always a minimum acceleration requirement at some finite Λ greater than

two. It has been found that a linear combination of the two requirements can be considered a critical maneuverability requirement, with the resulting miss rising sharply if the acceleration ability of the missile falls below this critical value.

The broken line on Fig. 8 shows the aggravating effect of any incidental or spurious feedback in the missile system which may finally limit the maximum λ obtainable.

A similar compromise governs the selection of the other major navigation parameter, the smoothing time. We have observed that if the flight time is sufficiently great in comparison to the smoothing time, the errors in the initial flight condition will have vanished, and insignificant miss will result from steady evasive maneuvers or incidental unbalances in the system. On the other hand,

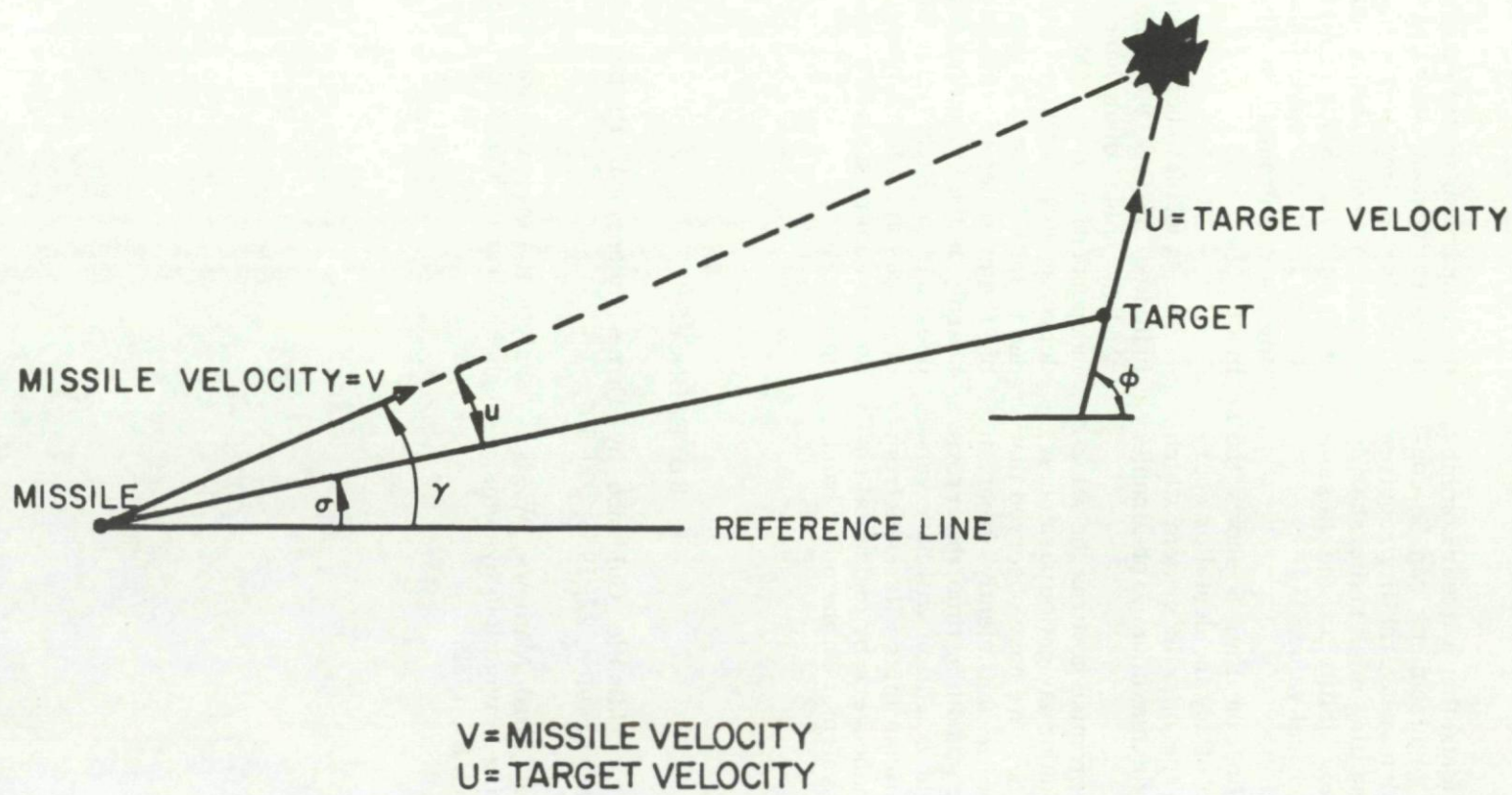
the dispersion of trajectories due to noise, tends to increase with decreased smoothing time, increasing very sharply when the critical maneuverability exceeds the ability of the missile. Since in any practical missile we are limited to some finite maximum flight time, a compromise in filter time must be made.

The tradeoff inherent in this situation is indicated in Fig. 9. The misses due to random noise disturbances are virtually independent of the flight time, but gradually decrease with increased filtering. The miss due to initial errors is dependent on the ratio of flight time to smoothing time, rising sharply as the smoothing time exceeds about one-tenth the flight time. Since the two types of misses are generally uncorrelated, a mean square summing gives a rational criterion for optimization.

REFERENCES

1. Adler, F. P., "Missile Guidance by Three-Dimensional Proportional Navigation," *Journal Applied Physics*, 27, 1956, 500-507.
2. Bennett, R. R., and Mathews, W. E., "Analytical Determination of Miss Distance for Linear Homing Navigation Systems," Hughes Aircraft Co., Technical Memorandum No. 260, March 31, 1952.

Fig. 1. Proportional navigation coordinate system.



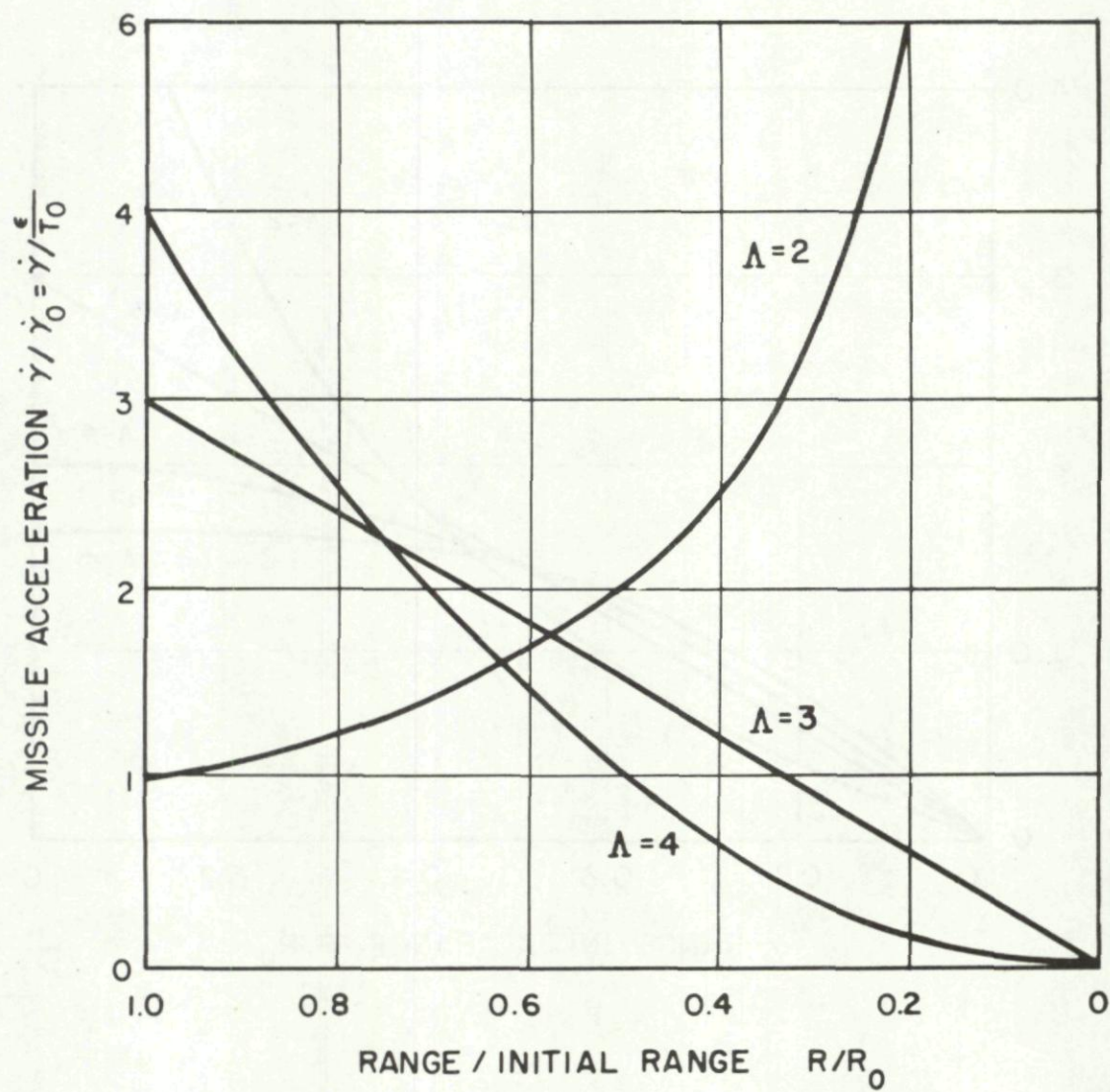


Fig. 2. Missile acceleration for initial heading error.

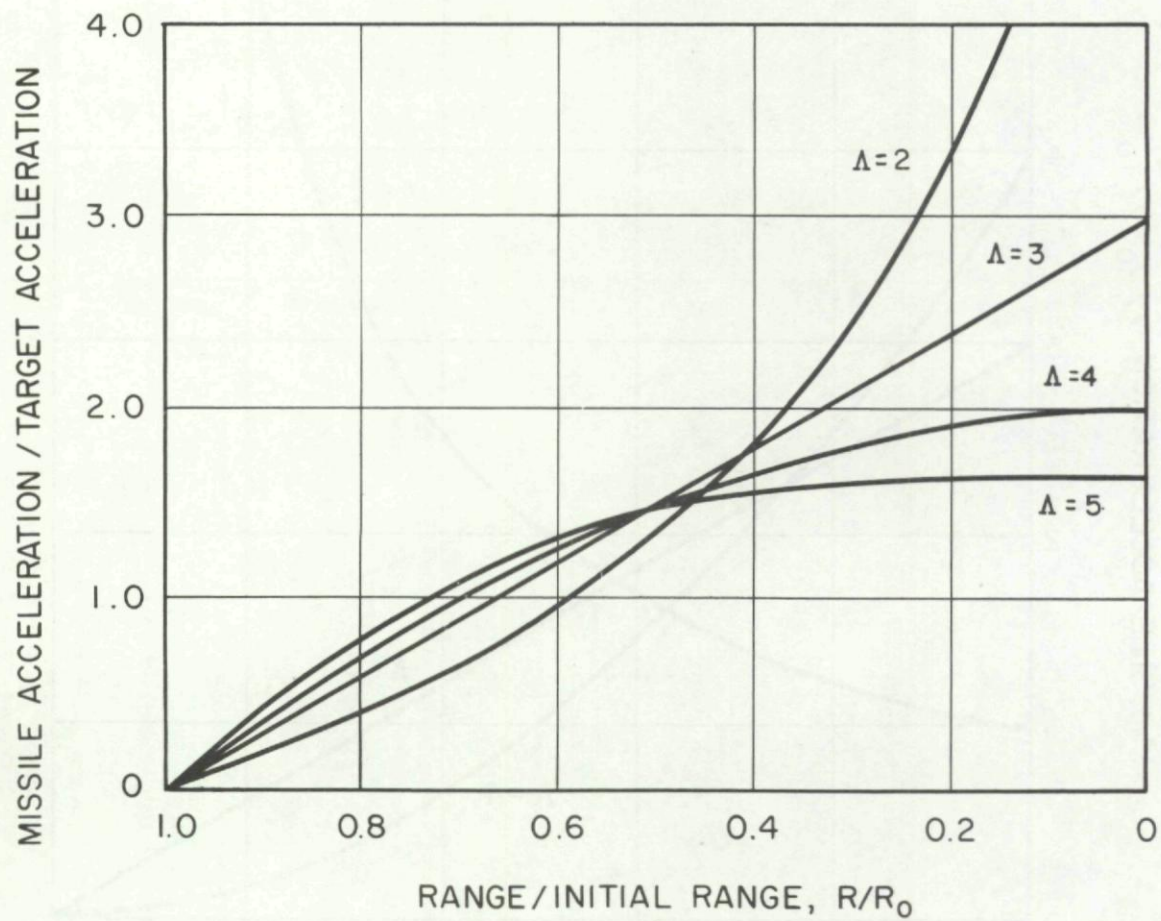


Fig. 3. Missile acceleration resulting from steady target acceleration.

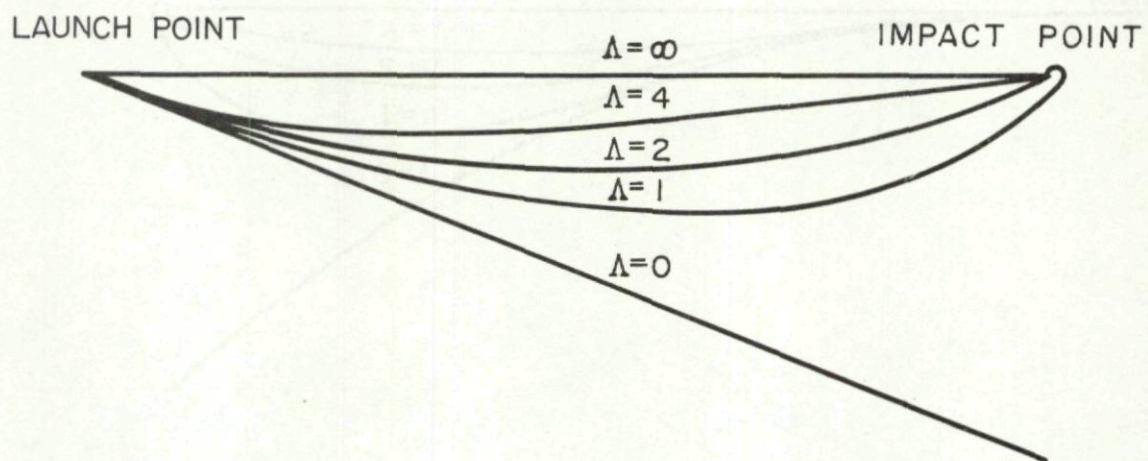


Fig. 4. Typical trajectories with initial error.

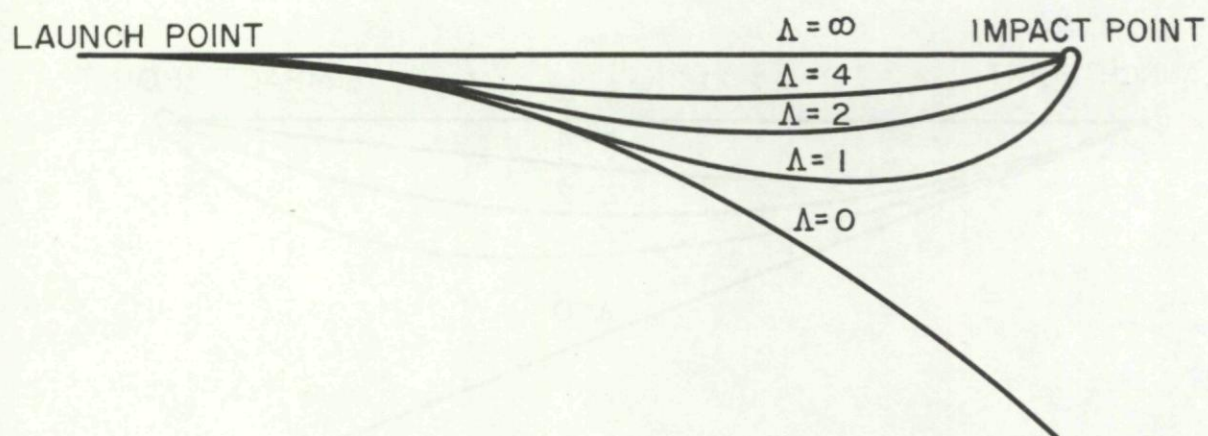
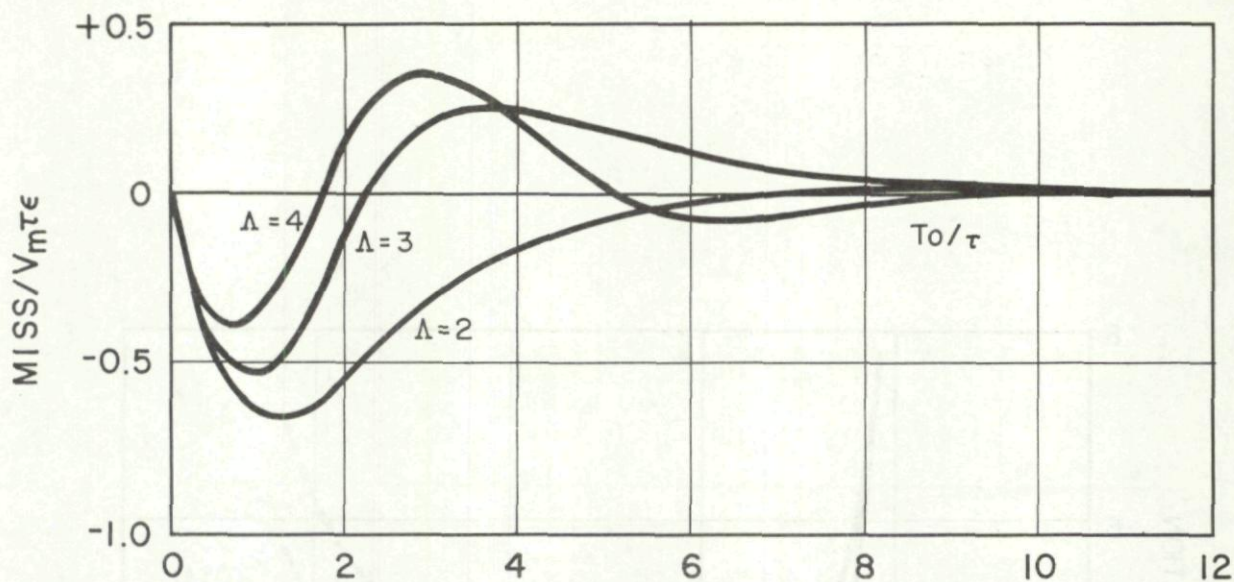
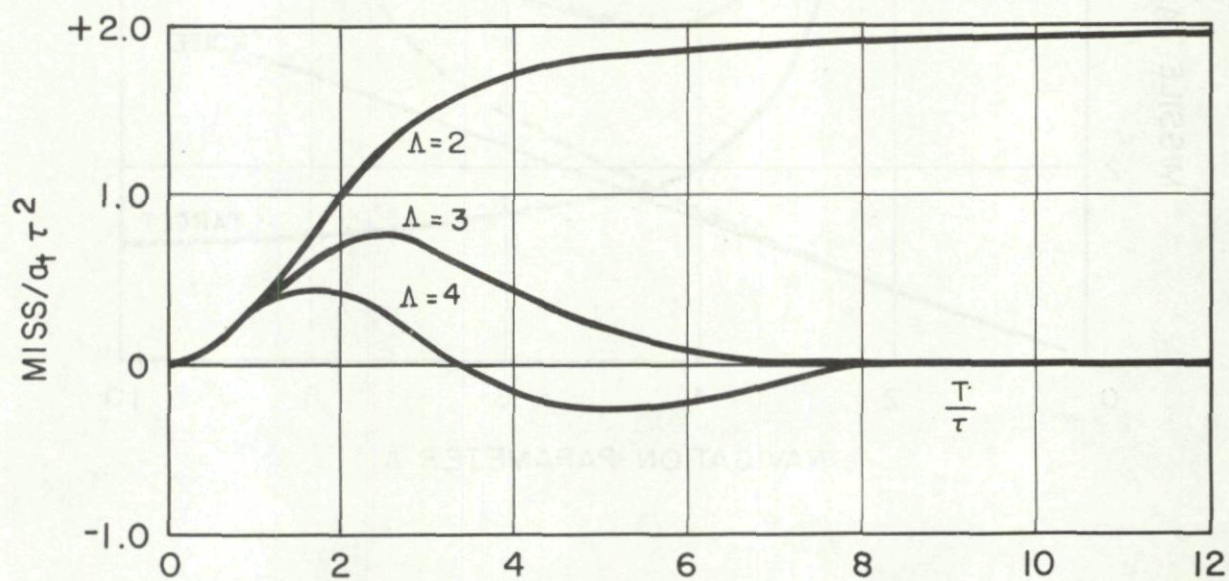


Fig. 5. Typical trajectories with gravity or $\dot{\gamma}$ drift.



(a) WEIGHTING FUNCTION FOR MISS DUE TO AIMING

$$\text{ERROR, } \epsilon, Z(p) = \left(1 + \frac{\tau}{2} p\right)^2$$



(b) MISS DUE TO STEP TARGET MANEUVER AT TIME-TO-GO, T .

Fig. 6. Weighting functions.

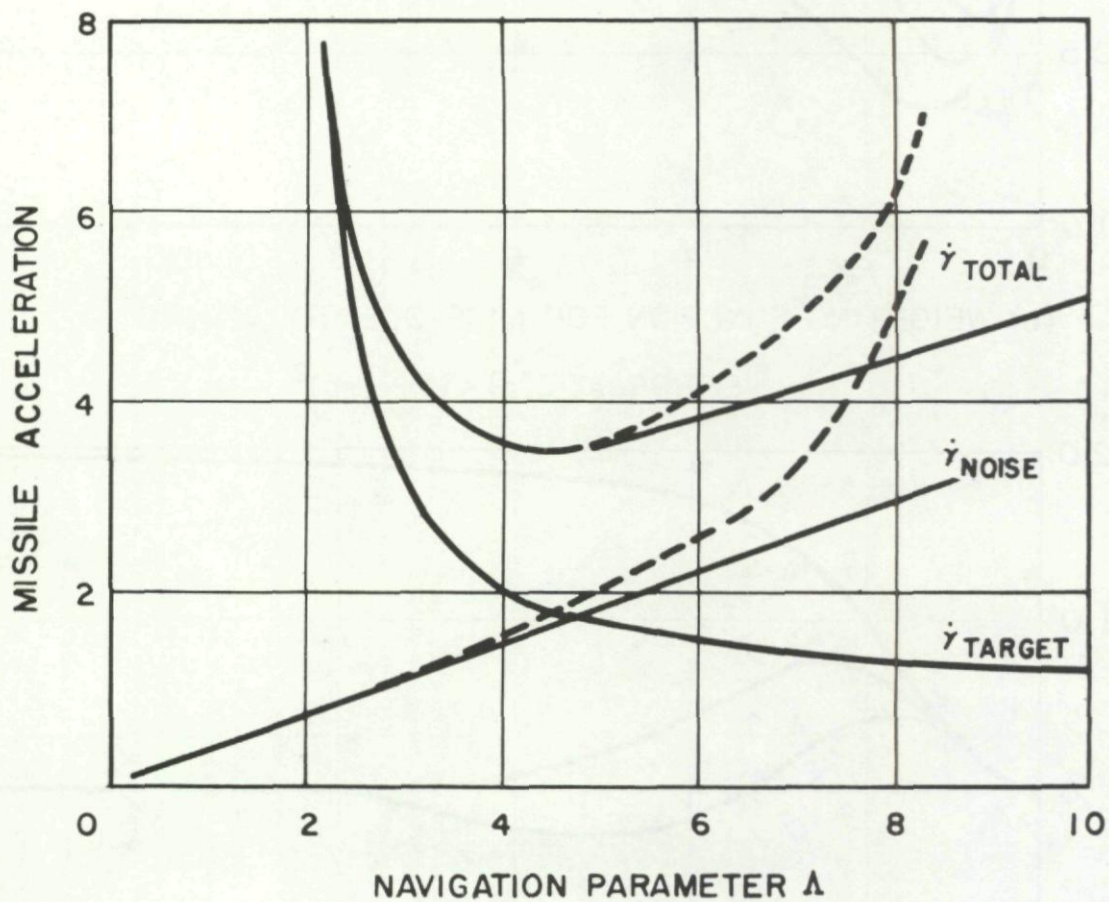


Fig. 7. Optimization of Λ by tradeoff of acceleration requirements.

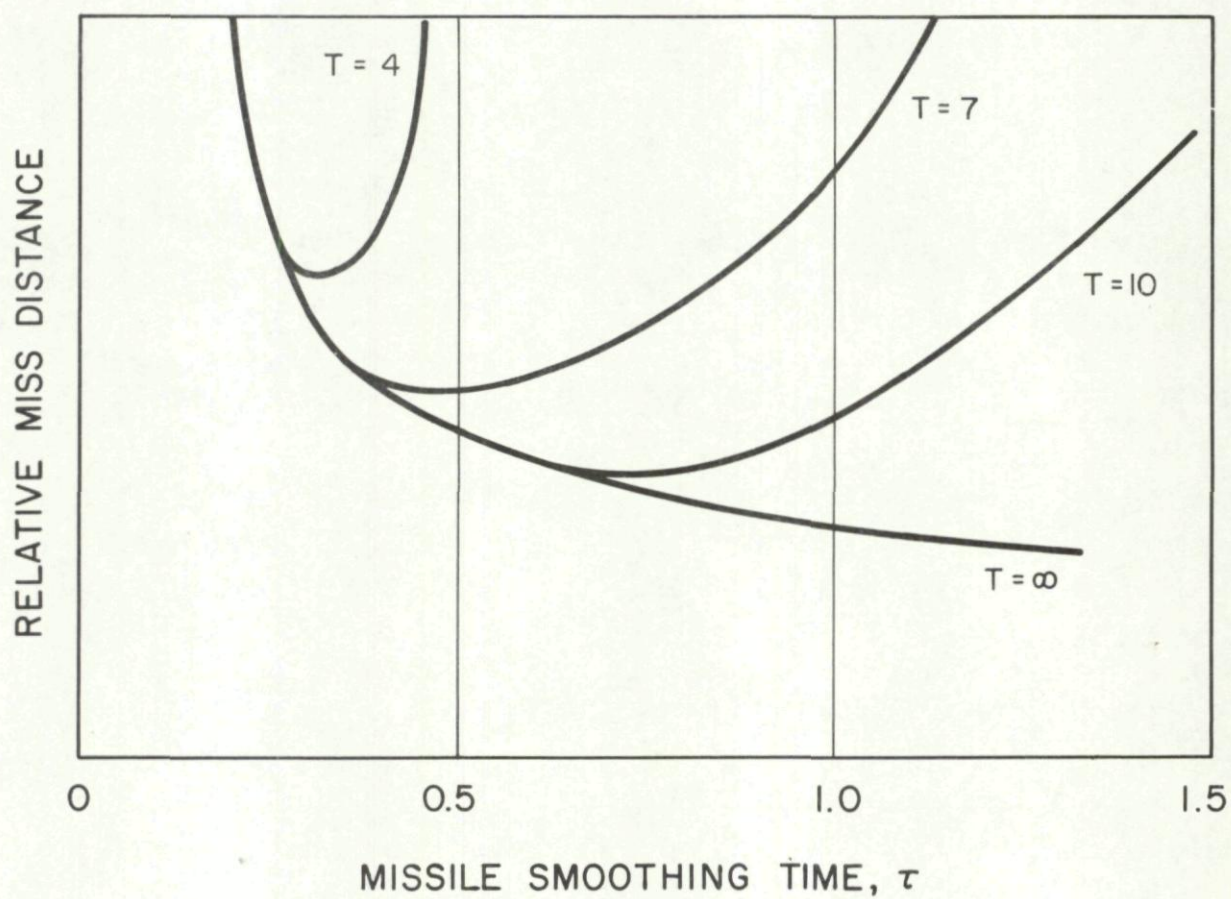


Fig. 8. Optimization of smoothing time by tradeoff between random dispersion and decay of initial errors.

PITFALLS IN MISSILE CONTROL

Robert L. Johnson*

SUMMARY

Unknown effects or problems are usually the ones which cause the most trouble in the realization of an adequately controlled missile. In an attempt to reduce this area of unknowns for others, this paper presents a number of problems which have been encountered in the design of past missiles. These include: static and dynamic aeroelasticity; structural feedbacks; cross-talk; aerodynamic, hydraulic, mechanical, and geometrical nonlinearities; overloads of various kinds; the adverse effects of noise, drift, and tolerances; and a mention of such operational problems as environmental and safety considerations. Since mathematical formulation usually follows directly from an understanding of the physical problem, the treatment is restricted to an explanation of the physics of typical examples in each case.

SOMMAIRE

Les phénomènes inconnus sont généralement ceux qui causent le plus d'ennuis dans la réalisation d'un missile contrôlé d'une manière adéquate. Dans le but de réduire le nombre de ces problèmes pour les autres, cette note présente un certain nombre de problèmes qui ont été rencontrés dans le passé dans l'étude des missiles. Ils comprennent: l'aéroélasticité statique et dynamique, les contre-réactions de structure; le mélange d'informations; les non-linéarités aérodynamiques, mécaniques, hydrauliques et géométriques; les surcharges de nature différente; les effets défavorables du bruit, de la dérive et des tolérances; et une mention des problèmes opérationnels tels que les considérations de sécurité et d'environnement. Bien que la mise sous forme mathématique suit habituellement directement la compréhension du problème physique cette note se réduit uniquement dans chaque cas à une explication des phénomènes physiques régissant des exemples typiques.

1. INTRODUCTION

An individual or an organization, in going from the desire to control a missile to the actuality of a controlled missile, passes through two general stages. The first is concerned with why control should work. It consists of learning the basic linear theories of servomechanisms and dynamics and their application in the form of basic components. The second stage revolves around the problems of why control does not work, at least initially. It consists of the

ground and flight tests which uncover the unknowns or pitfalls which were not included in the initial efforts.

These unknowns are not usually in themselves difficult to correct, once one realizes their existence. While they may complicate the design, or at least the effort in obtaining it, there are usually ways of handling the problems. This means that forewarning of typical problems should materially expedite the obtaining of adequate control. Prior

*Douglas Aircraft Company, Santa Monica, California.

consideration of known problems reduces the difficulties on a given missile to those which are peculiar to that missile.

The subject of this paper is a brief coverage of these known problems. General classes of problems covered are: linear problems, nonlinear problems, random effects, and operational problems. The treatment is restricted to an explanation of the physics involved in typical cases so that more examples may be given. Mathematical formulation usually follows directly from such a physical understanding. No attempt to be comprehensive has been made; only items which have caused specific difficulties in the author's experience are covered.

The advent of large scale analog and digital computing machines makes the distinction between linear and nonlinear problems somewhat unessential. It is based upon the "premachine" fact that linear problems could usually be solved and understood, while nonlinear problems could not. At present, the availability of computing machines for control analysis results in an understanding of the physical phenomena as the controlling factor. Ability to solve the mathematical equations describing the physics formerly was the controlling factor.

This does not mean that one can neglect learning the time-honored and basic techniques of linear analysis. Transfer functions, frequency characteristics, Nyquist diagrams, Bode plots, and root-locus techniques form an almost indispensable background for the controls designer. Neither does it mean that analysis of the problem, no matter how easy by machine techniques, is the end result. Many times it is possible to eliminate the source of the problem rather than live with it through analysis. Again, knowledge of the existence of problems is paramount, to which knowledge let us proceed.

2. LINEAR PROBLEMS

a. Aeroelasticity

Aeroelasticity is associated with the elastic deformation of the airframe under aerodynamic loads. Both static and dynamic effects must be considered.

An example of the static effect would be the deflection of a tail surface under the loads which it picks up at angles of attack. Such deflections change the loads which would otherwise be expected and lead to differences between actual and theoretical aerodynamic derivatives. In this case, the major effect would be on the lifting characteristics. Fig. 1 shows how the lift would be affected for a load aft of the surface torsional axis. The situation would be reversed for a load forward of the torsional axis. These changes in lift derivatives show up also as changes in moment derivatives. Both affect the dynamics of the airframe to be controlled.

Similar effects occur if actuating torque rods or the missile body deflect. Depending upon the particular airframe configuration involved, one must consider the effects of other structural deflections. It may well be that some parts will be critical in this regard, rather than in strength. In any event, the revised aerodynamic derivatives must be used in the automatic control analysis.

Such aeroelasticity, however, is not necessarily bad. One may employ it to modify aerodynamic characteristics as a function of load so as to minimize variations with air-speed and altitude. The effect upon flutter characteristics also must always be considered.

Dynamic aeroelasticity as used herein refers to the dynamic coupling of an aeroelastic phenomenon with the automatic control

system. More specifically, the end result of an automatic control system is the exertion of a control force of some kind. This force not only acts upon the aircraft considered as a rigid body but as an elastic body. If the elastic body deflects under this control force in a manner which can affect that force, a coupling or feedback exists which must be considered. Not only static but transient, i.e., dynamic, effects must be considered.

As an example, consider Fig. 2 which shows a block diagram of a control system containing a rate gyro and an accelerometer. A control surface deflection will not only excite the missile as a rigid body but it will impose forces on the missile which deflect it, i.e., excite it, as an elastic body. The missile in deflecting will move the rate gyro and the accelerometer and the resulting signals will, through the position servo, result in further control surface deflections. The element "structural dynamics" is not a simple one because all missiles have several modes of vibration in which they can deflect. Usually one finds that the most important modes are the first, second, etc., bending modes. Fig. 3 indicates the shape of the first three of these modes. Fig. 4 shows how, on the first mode, the accelerometer and rate gyro in the example interpret this bending motion.

Associated with each mode is not only a shape but a natural frequency and a damping ratio. Therefore, one can plot the frequency behavior of a missile as a transfer function relating output motion to input force or control deflection at a given point. The motion in our example may be either the vertical motion at the accelerometer location or the angular motion at the rate gyro location. Fig. 5 shows a typical plot where, in the interests of simplicity, the phase characteristics are not shown. They, as well as the amplitude characteristics, may be determined either experimentally or theoretically.

With the structural dynamics known, it is a simple matter to include the extra feedback loop in the equations governing the system. Stability may be determined as before and, if necessary, steps taken to obtain it. These steps may consist of electrical or mechanical filters or placing the instruments so that they are insensitive to the bending motions.

b. Structural Feedbacks

There may be other feedback paths through the structure not associated directly with airloads and hence not classed with the aeroelastic phenomena. A good example is the vibration induced by the inertia loads imposed on the structure by high performance servo systems. The rapid accelerations of which these systems are capable result in both translational and rotational excitation being applied to the airframe. If there are local resonances in any of the control instruments or their mounts, in control valves or in servoamplifiers, instability may result. The problem of instrument mounts should be mentioned separately. Too often the rigidity of the mounting structure is not given adequate attention. It is quite clear that if the natural frequency of the mount (including the weight of the instrument) is below that of the instrument, the mount frequency will be the controlling dynamic phenomenon. The additional unsuspected phase lags which result may lead to dynamic instability around or above the mount natural frequency.

Another resonance which might be excited by wing inertia loads, motor roughness, or other excitation feeding through aircraft structure is internal instrument resonance. Examples are potentiometer brushes, pressure gage diaphragms, etc. All of these effects can be handled in a linear analysis by the same technique as presented for the aeroelastic phenomena. However, a better approach is to eliminate the resonance rather than design around it.

c. Cross-talk

Cross-talk is the appearance in one control axis of signals which are intended for or result from actions in another control axis. Consider a vertical signal vector in earth coordinates being resolved into cruciform missile coordinates with one set of wings level. With the wings in fact level, the resolution results in only an up maneuver. However, if the missile is rolled somewhat, a horizontal maneuver also results, as shown in Fig. 6. The one signal has "cross-talked" into the other axis. The change in the intended direction is usually small. In closed loop systems, cross-talk errors will tend to be corrected but there are limits to how large this coupling can be.

Cross-talk arises from a number of different effects. Examples are:

- (1) Instruments are not mounted in the proper direction so that they are sensitive to motions in other axes.
- (2) Servo valve spools are sensitive to accelerations in another axis from that which they control.
- (3) Rate gyros are becoming sensitive in another axis as they deflect while measuring rates about the intended axis.
- (4) Rate gyros are acting like angular accelerometers about the gimbal axis.
- (5) Three-dimensional coupling due to the nonorthogonality of free gyro gimbals.
- (6) Aerodynamic coupling resulting from steering angles of attack and control surface angles.

When linear or linearized, these cross-talk effects may be handled by analytic techniques. Usually they are nonlinear and must be considered by other methods.

It should be noted that the above list only provides examples and is not nearly complete. Nor can it be complete because each system will have its own problems which can only be uncovered by the designer as he considers the effects of departures from the desired state of affairs.

3. NONLINEAR PROBLEMS

As previously noted, nonlinear problems have been considered separately only because the techniques for handling them are different. The equations of motion are still written in the same way although many times the nonlinearities are such that the equations are difficult to formulate. However, they may no longer be handled through usual operational calculus methods and previously mentioned graphical techniques. Solution of these nonlinear problems usually requires analog or digital computing machine techniques. Linearization about some quiescent operating point is often employed with good results if interpreted in the light of a restricted solution.

a. Aerodynamic Nonlinearities

The aerodynamic characteristics of the airframe to be controlled are clearly of major importance to the control system designer. In fact, as indicated in the paper by Perry, the airframe characteristics are so important that the configuration must not be designed and presented to the controls man as a "fait accompli." Consideration of control requirements must help establish the configuration if a balanced design is to be obtained. For example, some of the bending

system. More specifically, the end result of an automatic control system is the exertion of a control force of some kind. This force not only acts upon the aircraft considered as a rigid body but as an elastic body. If the elastic body deflects under this control force in a manner which can affect that force, a coupling or feedback exists which must be considered. Not only static but transient, i.e., dynamic, effects must be considered.

As an example, consider Fig. 2 which shows a block diagram of a control system containing a rate gyro and an accelerometer. A control surface deflection will not only excite the missile as a rigid body but it will impose forces on the missile which deflect it, i.e., excite it, as an elastic body. The missile in deflecting will move the rate gyro and the accelerometer and the resulting signals will, through the position servo, result in further control surface deflections. The element "structural dynamics" is not a simple one because all missiles have several modes of vibration in which they can deflect. Usually one finds that the most important modes are the first, second, etc., bending modes. Fig. 3 indicates the shape of the first three of these modes. Fig. 4 shows how, on the first mode, the accelerometer and rate gyro in the example interpret this bending motion.

Associated with each mode is not only a shape but a natural frequency and a damping ratio. Therefore, one can plot the frequency behavior of a missile as a transfer function relating output motion to input force or control deflection at a given point. The motion in our example may be either the vertical motion at the accelerometer location or the angular motion at the rate gyro location. Fig. 5 shows a typical plot where, in the interests of simplicity, the phase characteristics are not shown. They, as well as the amplitude characteristics, may be determined either experimentally or theoretically.

With the structural dynamics known, it is a simple matter to include the extra feedback loop in the equations governing the system. Stability may be determined as before and, if necessary, steps taken to obtain it. These steps may consist of electrical or mechanical filters or placing the instruments so that they are insensitive to the bending motions.

b. Structural Feedbacks

There may be other feedback paths through the structure not associated directly with airloads and hence not classed with the aeroelastic phenomena. A good example is the vibration induced by the inertia loads imposed on the structure by high performance servo systems. The rapid accelerations of which these systems are capable result in both translational and rotational excitation being applied to the airframe. If there are local resonances in any of the control instruments or their mounts, in control valves or in servoamplifiers, instability may result. The problem of instrument mounts should be mentioned separately. Too often the rigidity of the mounting structure is not given adequate attention. It is quite clear that if the natural frequency of the mount (including the weight of the instrument) is below that of the instrument, the mount frequency will be the controlling dynamic phenomenon. The additional unsuspected phase lags which result may lead to dynamic instability around or above the mount natural frequency.

Another resonance which might be excited by wing inertia loads, motor roughness, or other excitation feeding through aircraft structure is internal instrument resonance. Examples are potentiometer brushes, pressure gage diaphragms, etc. All of these effects can be handled in a linear analysis by the same technique as presented for the aeroelastic phenomena. However, a better approach is to eliminate the resonance rather than design around it.

c. Cross-talk

Cross-talk is the appearance in one control axis of signals which are intended for or result from actions in another control axis. Consider a vertical signal vector in earth coordinates being resolved into cruciform missile coordinates with one set of wings level. With the wings in fact level, the resolution results in only an up maneuver. However, if the missile is rolled somewhat, a horizontal maneuver also results, as shown in Fig. 6. The one signal has "cross-talked" into the other axis. The change in the intended direction is usually small. In closed loop systems, cross-talk errors will tend to be corrected but there are limits to how large this coupling can be.

Cross-talk arises from a number of different effects. Examples are:

- (1) Instruments are not mounted in the proper direction so that they are sensitive to motions in other axes.
- (2) Servo valve spools are sensitive to accelerations in another axis from that which they control.
- (3) Rate gyros are becoming sensitive in another axis as they deflect while measuring rates about the intended axis.
- (4) Rate gyros are acting like angular accelerometers about the gimbal axis.
- (5) Three-dimensional coupling due to the nonorthogonality of free gyro gimbals.
- (6) Aerodynamic coupling resulting from steering angles of attack and control surface angles.

When linear or linearized, these cross-talk effects may be handled by analytic techniques. Usually they are nonlinear and must be considered by other methods.

It should be noted that the above list only provides examples and is not nearly complete. Nor can it be complete because each system will have its own problems which can only be uncovered by the designer as he considers the effects of departures from the desired state of affairs.

3. NONLINEAR PROBLEMS

As previously noted, nonlinear problems have been considered separately only because the techniques for handling them are different. The equations of motion are still written in the same way although many times the nonlinearities are such that the equations are difficult to formulate. However, they may no longer be handled through usual operational calculus methods and previously mentioned graphical techniques. Solution of these nonlinear problems usually requires analog or digital computing machine techniques. Linearization about some quiescent operating point is often employed with good results if interpreted in the light of a restricted solution.

a. Aerodynamic Nonlinearities

The aerodynamic characteristics of the airframe to be controlled are clearly of major importance to the control system designer. In fact, as indicated in the paper by Perry, the airframe characteristics are so important that the configuration must not be designed and presented to the controls man as a "fait accompli." Consideration of control requirements must help establish the configuration if a balanced design is to be obtained. For example, some of the bending

difficulties mentioned earlier can be eased by proper choice of length to diameter ratio. Also, many of the aerodynamic nonlinearities mentioned below can be eliminated by proper choice of configuration. Such nonlinearities are not necessarily fatal but their effects must certainly be considered.

One type of nonlinearity is that shown in Fig. 7 which occurs near zero angle of attack. This type is an actual static instability at small angles of attack. Depending upon the type of control feedback employed, this can lead to low amplitude oscillations wasteful in control energy and in kinetic energy, i.e., drag. Another type, shown in Fig. 8, represents a loss of restoring moment at high angles of attack. This can lead to tumbling the missile if it overshoots or is commanded into the poor control region. Stability changes like this are usually associated with viscous crossflow effects. Hence, they may be a severe function of bank angle which can be ruinous to an agile missile.

Another type of aerodynamic nonlinearity, which could also be mentioned under the cross-talk heading, is that of pitch/yaw/roll interactions. These result in rolling moments due to steering, and in steering moments, due to rolling. Loss of stability at high angles of attack at different bank angles as indicated in Fig. 8 is an example of the latter. Fig. 9 shows typical rolling moment curves for various steering conditions. Care must be taken that these rolling moments do not overpower the roll control system, either statically or dynamically. Fig. 10 shows the dynamic response of a missile attempting to hold a zero roll position under the rolling moments imposed by noisy steering commands. The dashed curve shows the response before the real magnitudes of the rolling moments were appreciated. The solid curve shows the greatly improved and entirely adequate behavior after dynamic pressure sensitive gain changing circuits were added.

b. Hydraulic Nonlinearities

Perhaps the most annoying nonlinearity associated with hydraulics is the tendency for broken lines to spray oil on everyone except hydraulic engineers. This tendency not being subject to analysis, let us pass on to typical hydraulic nonlinearities.

Flow versus plunger displacement in control valves is apt to be nonlinear because of land overlap and velocity squared pressure drop across orifices. Fig. 11 illustrates the basic characteristic. Also indicated on Fig. 11 by the dotted lines are the quasi-linear gain change effects which result for different amplitudes of motion. These provide a reasonable way of determining whether such gain changes would be troublesome. For example, interpreting the Nyquist diagram of Fig. 12 shows that a gain reduction of 3.5 would probably result in instability at the frequency indicated as f_2 . This would be a limited low-amplitude oscillation because higher amplitudes would increase the gain and provide stability. Such oscillations can be troublesome in terms of loss of hydraulic oil, increase in drag, decreased accuracy, etc.

Actually, the control valve characteristic of final importance is not from plunger displacement to flow, but from control current (or voltage) to flow. This introduces other nonlinearities such as control current to plunger force and plunger position to plunger force. This latter is associated with the axial forces imposed on sliding valve plungers as they divert the flow. Experimental determination of both the component and overall nonlinearities is quite feasible if one is alerted to look for them. Compressibility of oil, either from its own bulk modules or from entrapped air, introduces nonlinearities which are dependent upon the various volumes throughout the system. Closely associated

with this is the apparent compressibility introduced by expansion of actuating cylinders and lines under pressure loads. In general, these effects appear as phase lags and unwanted resonances.

c. Mechanical Nonlinearities

The usual mechanical nonlinearities of dead spot, dry friction, preloads, etc., enter most engineers' thinking at once. One which is less obvious is the step by step output voltage of a wire wound potentiometer as the brush is moved smoothly from one end to the other. In high gain servo systems, these step functions can excite resonances or lead to low amplitude oscillations as the system oscillates from turn to turn. Fig. 13 shows the results of an analog study on control systems having successively coarser resolution. The continued oscillations were also encountered in flight occasionally.

d. Geometrical Nonlinearities

Many nonlinearities arise purely from geometrical considerations while operating over large angles. For example, the conventional aerodynamic equations assuming small angles, so that the cosine equals one and the sine equals the angle, may not be valid for missiles operating at high angles of attack. In handling three-dimensional problems, the actual direction cosines or Euler angle transformations must be used which are nonlinear. The free gyro gimbal nonorthogonality mentioned earlier is a case in point. Many missile trajectories entail large changes in the velocity vector which involve three-dimensional problems.

A less obvious geometrical nonlinear problem is that involved in the differentiation of periodic functions. For example, rather than install a rate gyro to provide damping in

a control system, one sometimes differentiates the position signal. Usually one tacitly assumes that the position output is linear, and it is usually for the working range. However, for larger angles, the position pickoff is usually periodic, putting out either a sine wave or a triangular wave (Fig. 14).

Within the range ± 45 degrees, both the approximate magnitude and the sign of $d(\text{Angle})/dt = d(\text{Voltage})/dt$ are correct. Between ± 90 degrees, the sign is at least correct. But beyond 90 degrees, the sign of $d(\text{Voltage})/dt$ has reversed, even though $d(\text{Angle})/dt$ is the same. This means that in this region the damping voltage has reversed and is now undamping the system. A cyclic instability, although not sinusoidal, is usually the result when one gets past the 90 degree points. For the triangular wave, the system behaves linearly out to 90 degrees. For the sine wave, the damping drops off as one approaches 90 degrees because of the reduced slope.

e. Saturating

A frequent type of system nonlinearity is that associated with the saturating, i.e., reaching the limit of a given quantity, even though within the operating range the quantity may be linear. Typical examples are:

- (1) The saturating of instruments such as accelerometers and rate gyros when their operating range is exceeded. Stops are usually present to avoid damaging the instruments under such conditions.
- (2) Control surface limits in either position, velocity, or acceleration. The former is a function of the airframe design, while the latter two are actuator considerations.

- (3) Intentional limiters are often placed in control systems so that damaging maneuvers will not be requested. These may be command limiters on input voltage or they may be torque limitations on the control surface actuators.

f. General Comments

The above represent typical physical phenomena, the effects of which must be considered in any given design. However, one need not live with all of them. If their ill effects are realized, some may be avoided by suitable design selection. This is certainly true of overloads in instrument ranges, proper choice of potentiometer resolution, etc. Sometimes the ill effects of several may cancel. For example, a dead spot in the system may stop an oscillation which would be present due to potentiometer resolution. If the adverse effect of the dead spot on accuracy can be tolerated, this may be an acceptable solution. In general, however, this is probably risky business, especially if the nonlinearities are not stable with time and component changes. The point of these comments is that realizing a problem's existence either permits avoidance, or the best compromise with it.

4. RANDOM EFFECTS

Previously mentioned problems are of a fixed nature, that is, the problem can be described once and for all and a solution obtained. However, there are problems arising from random effects which must be considered on a statistical basis. The three to be considered here are noise, drift, and tolerances. One cannot eliminate these problems by special design because they are always present. Rather, one designs so that their effect is tolerable.

a. Noise

Noise is the presence somewhere in the loop, usually on the intelligence signals, of unwanted random disturbances. Examples are the scintillation of radar signals, the fading and static on radio signals, power supply noise due to vibrator noise or some governor action, potentiometer scratch, etc. All these affect control in the same general ways:

- (1) Introduction of errors due to saturation and rectification as illustrated in Fig. 15. The presence in the system of limit of some kind, structural, electrical, etc., clips the top off the noise as shown by the shaded portions. Since the bottom is not clipped, the result is an error between the reported apparent average and the actual average.

- (2) The excessive use of some form of system power, either electrical, hydraulic, propulsive, or aerodynamic (drag). There really is no need to follow all this noise with the missile because the target or desired reference certainly isn't doing it. Trying to do so is wasteful of energy.

The usual method of reducing the effects of noise when the source cannot be eliminated is to provide filtering of some kind. Various kinds of notch or low-pass filters may be employed.

b. Drift

Drift is the inability of circuits to maintain calibration about their DC level. It may be considered to be a very low frequency noise but is mentioned separately because it affects designs in a different manner. One does not filter drift but operates at DC levels which are such that drift is a small value of the

controlled quantity. For example, a reasonable value for drift on the first grid of a DC amplifier is 50 millivolts. Clearly then, one must not operate at input signal levels where 50 millivolts is a significant part of the input to the amplifier. Such things as temperature, time, vibration, tube aging, etc., all effect drift. There are ways of stabilizing amplifiers with respect to drift such as chopper stabilization. However, they are not as suitable for airborne use as for ground installations.

c. Tolerances

When one has selected certain circuit, mechanical, and hydraulic values for use in a certain design, it is a great temptation to demand that the equipment be manufactured to those values. Unfortunately, no manufacturing organization can function in this manner because the rejection rate would be prohibitive. There must be some tolerance on every dimension, circuit value, gain, etc. Clearly, the final design must be capable of manufacture and, hence, the article cannot be considered designed until it can be manufactured within the limitations of existing methods. Therefore, in establishing circuits, one must allow for 1 percent resistors to become 5 percent with age and for amplifier gain to change ± 3 db. Structural alignments must be within tooling limits. Their effects on aerodynamic derivatives must be considered and, if unacceptable, either the basic design changed or better manufacturing methods conceived. The design-manufacturing cycle is a closed loop, too, and one part must be allowed to feed back to the other.

5. OPERATIONAL PROBLEMS

There are still items which must be considered before a design can be considered complete. Once manufactured, it will be in

operational use and it must operate in environments which are foreign to the engineering department. Furthermore, one must consider the consequences of its not working.

a. Environment

The environment within which the system must operate may be considered to consist of climatic, dynamic, and educational environments. By climatic environments we mean temperature, humidity, sand and dust, fungus, rain, snow, mud, etc. To how many of these items will the system components be subjected? If subjected to certain conditions, they must either operate or be protected so the conditions won't affect them.

The dynamic environment may be considered to be the shock, vibration, and steady-state accelerations to which system components are subjected. This must include shipment, storage, and final use conditions.

The educational environment is the intelligence and training level of the people who will be operating and maintaining the equipment. It does little good to put out the best possible system if it requires a team of doctors of engineering to operate it. Simplicity and reliability combined with a little human engineering are the keynotes. Here again, it is impossible to put down general numbers or equations. The above environments must be considered and applied with suitable margins of safety to each specific case.

b. Safety

All of this paper thus far has been devoted to designing the system so it will work. But what if it doesn't work occasionally once it gets into service? The consequences of such failure must be considered and, perhaps, the design affected accordingly. For example, what if the servoamplifier of an air launched

guided missile fails at launch and drives the control surfaces hard over. Clearly the missile should not strike the launching aircraft. Perhaps this requires a time delay before the control surfaces can move or break links to allow the control surface to center under excessive load. Other points where safety considerations should affect the design will undoubtedly arise in any given missile system.

6. CONCLUSION

In a sense, this paper has had an opportunity to present little more than a list of past difficulties. Proper employment should prevent their becoming future difficulties, at least without warning. Unfortunately, there is no complete list of the unknowns of missile control. Future efforts would then be so easy.

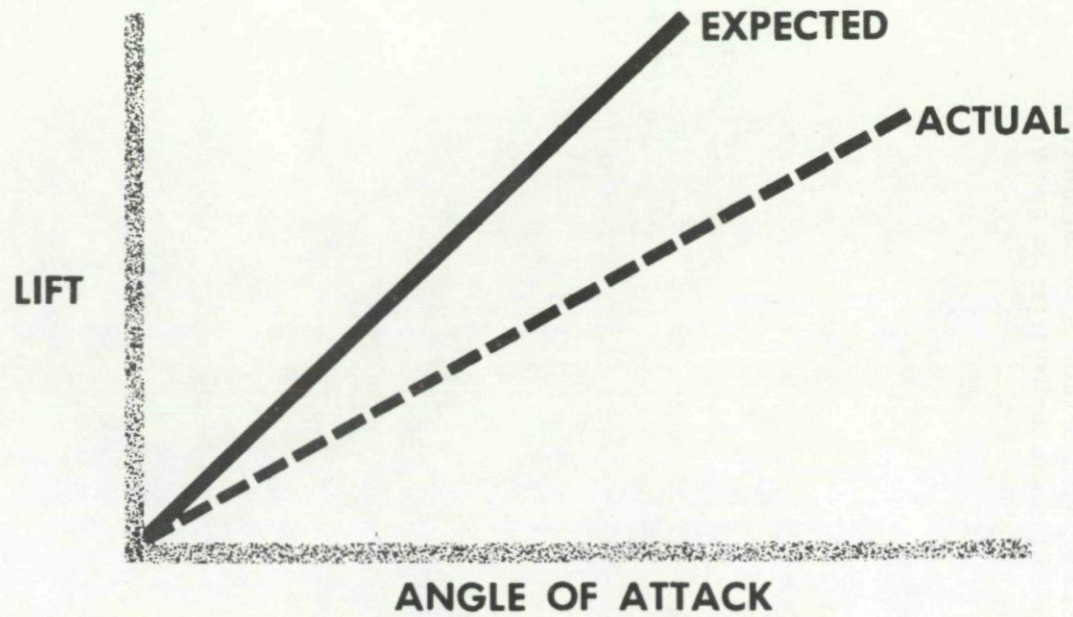
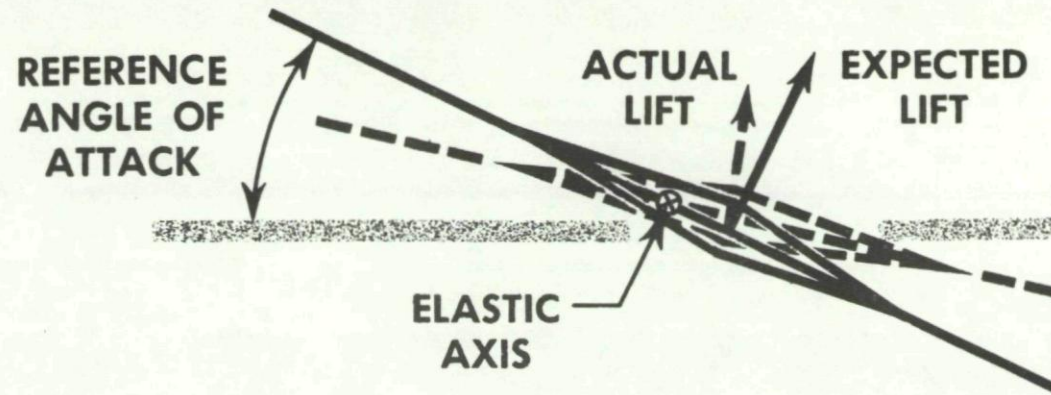


Fig. 1. Static aeroelasticity.

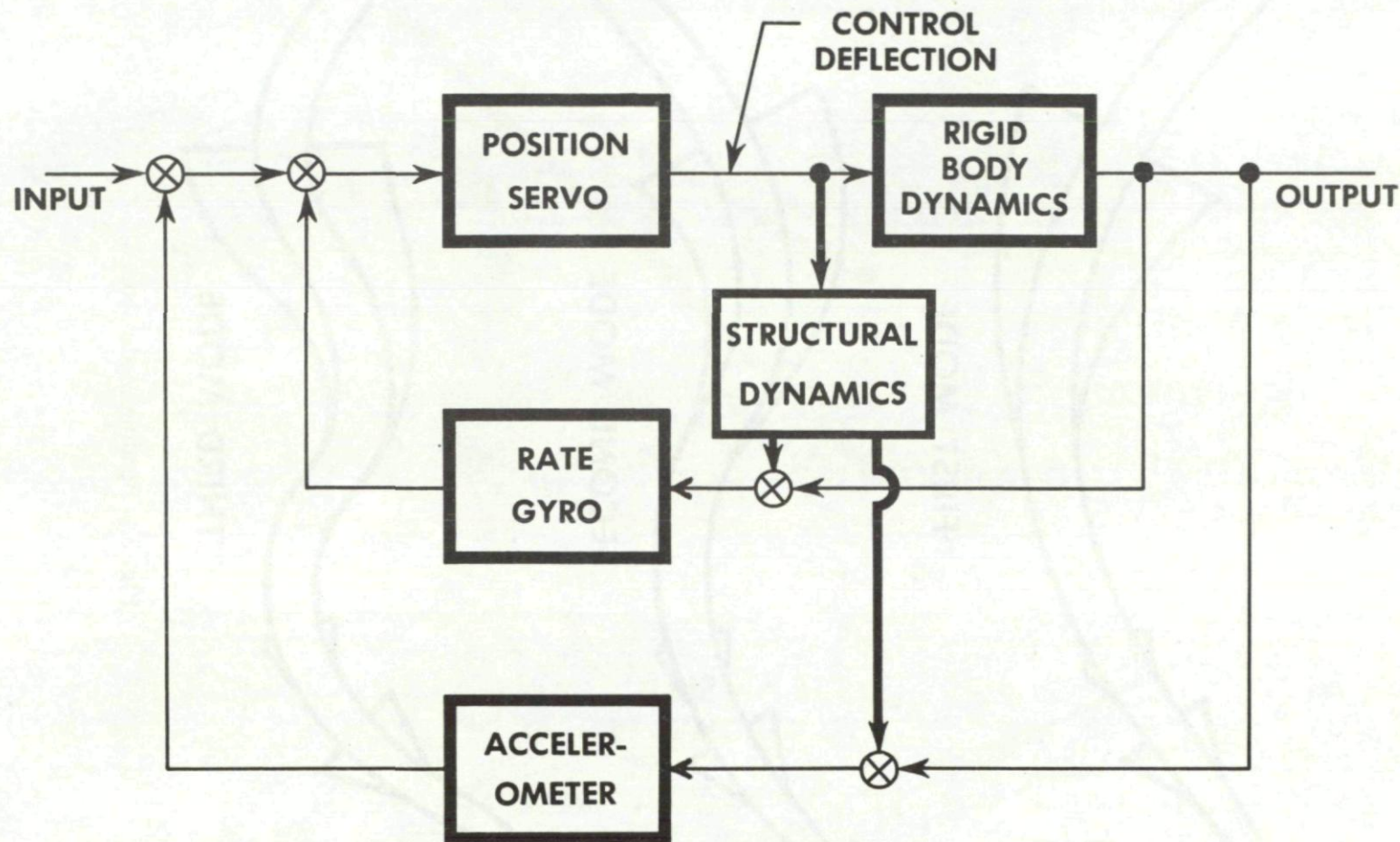
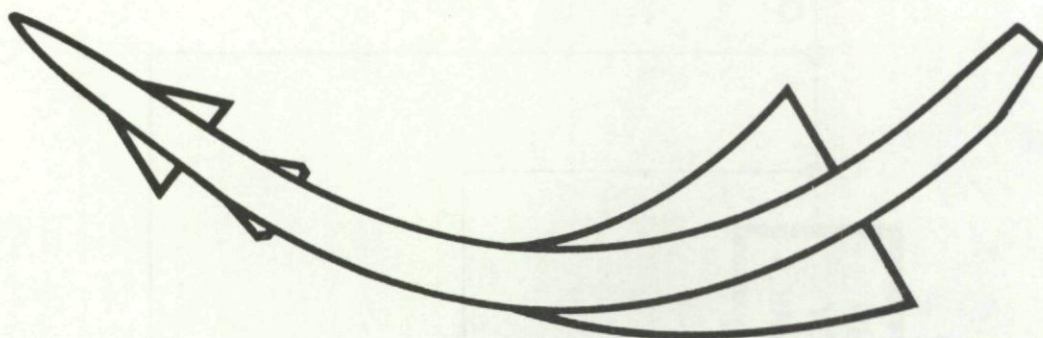
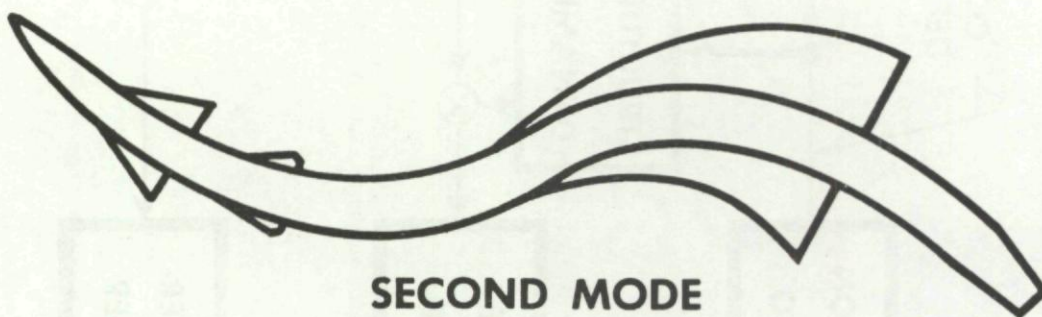


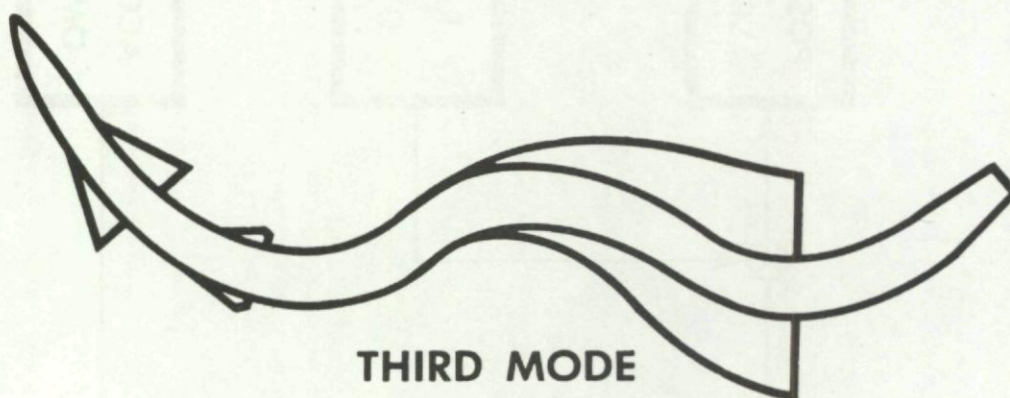
Fig. 2. Dynamic aeroelasticity feedback.



FIRST MODE



SECOND MODE



THIRD MODE

Fig. 3. Bending mode shapes.

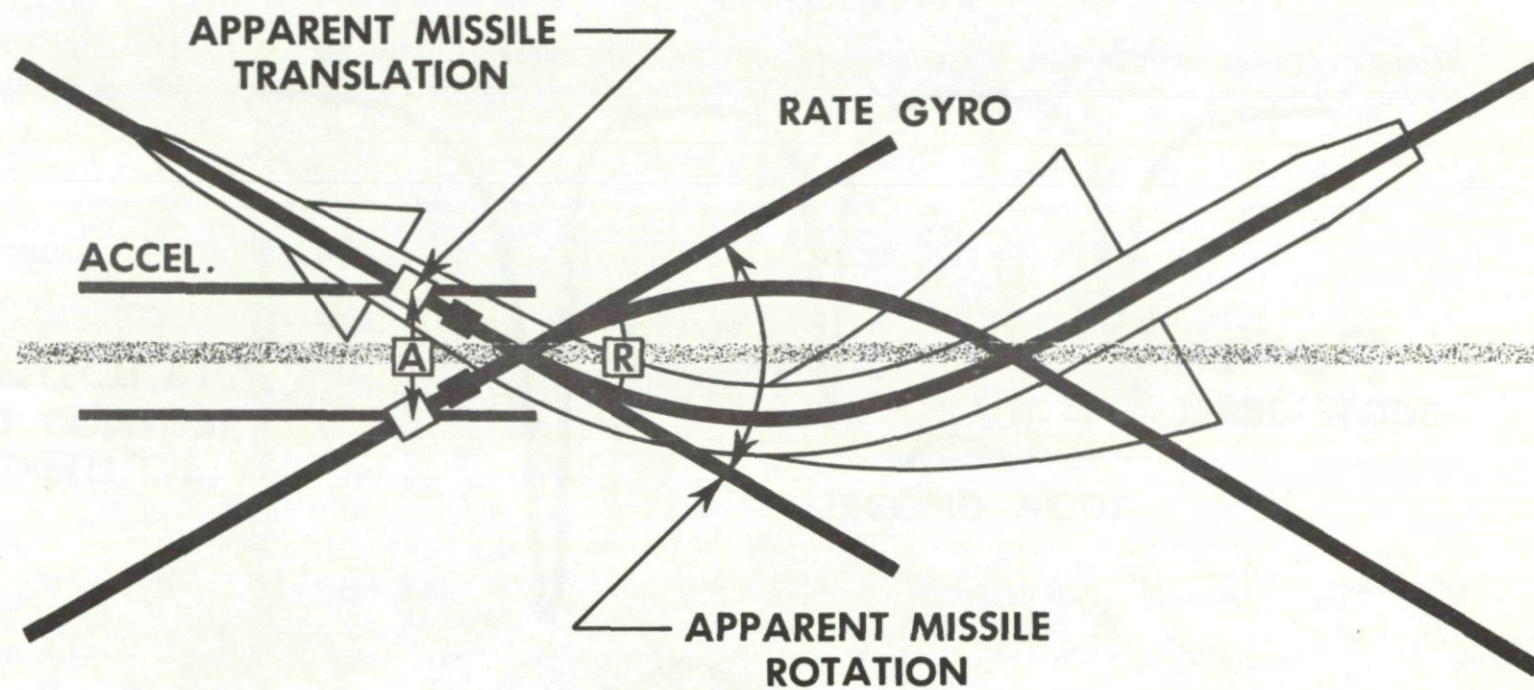


Fig. 4. Instrument interpretation of bending.

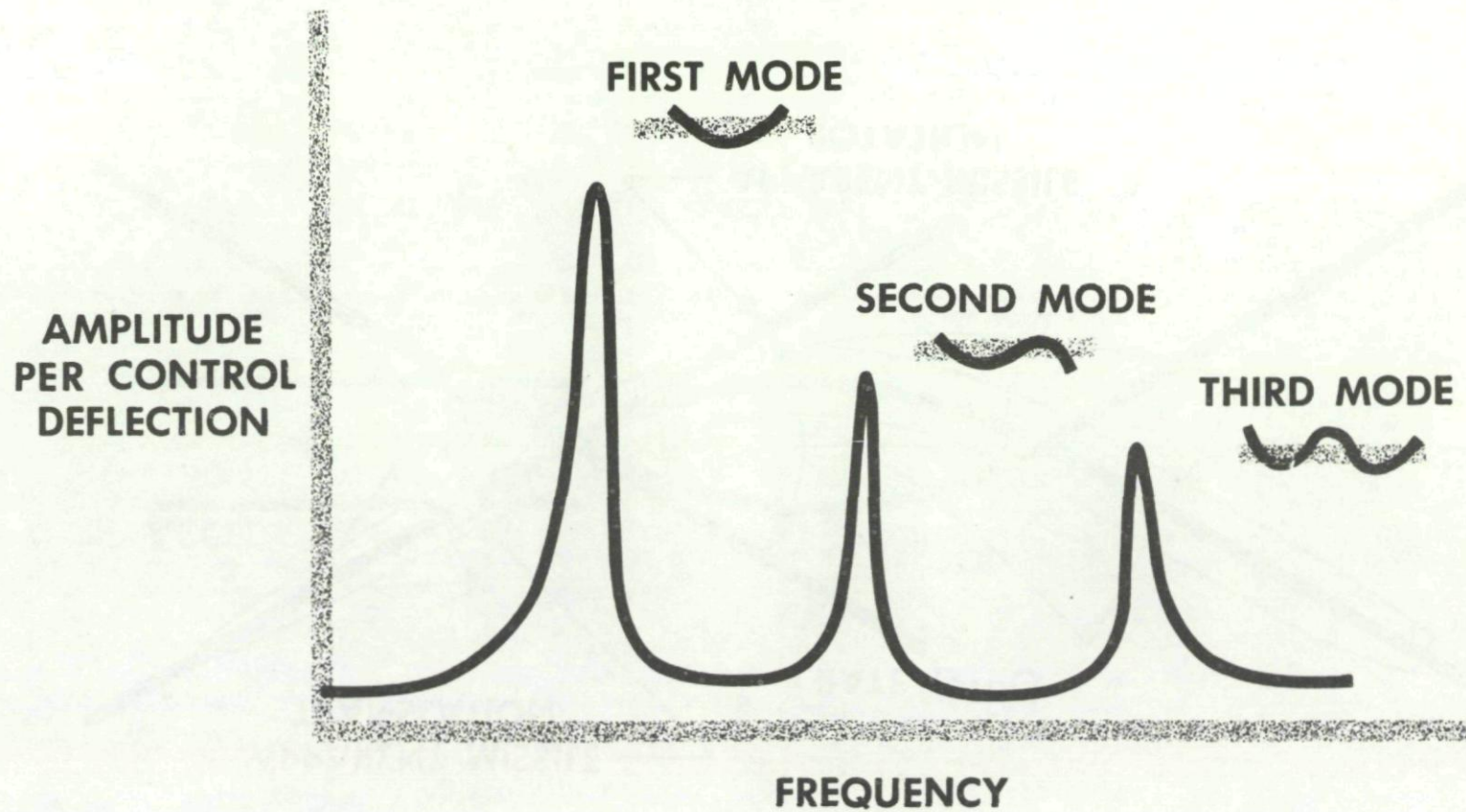


Fig. 5. Typical structural resonances.

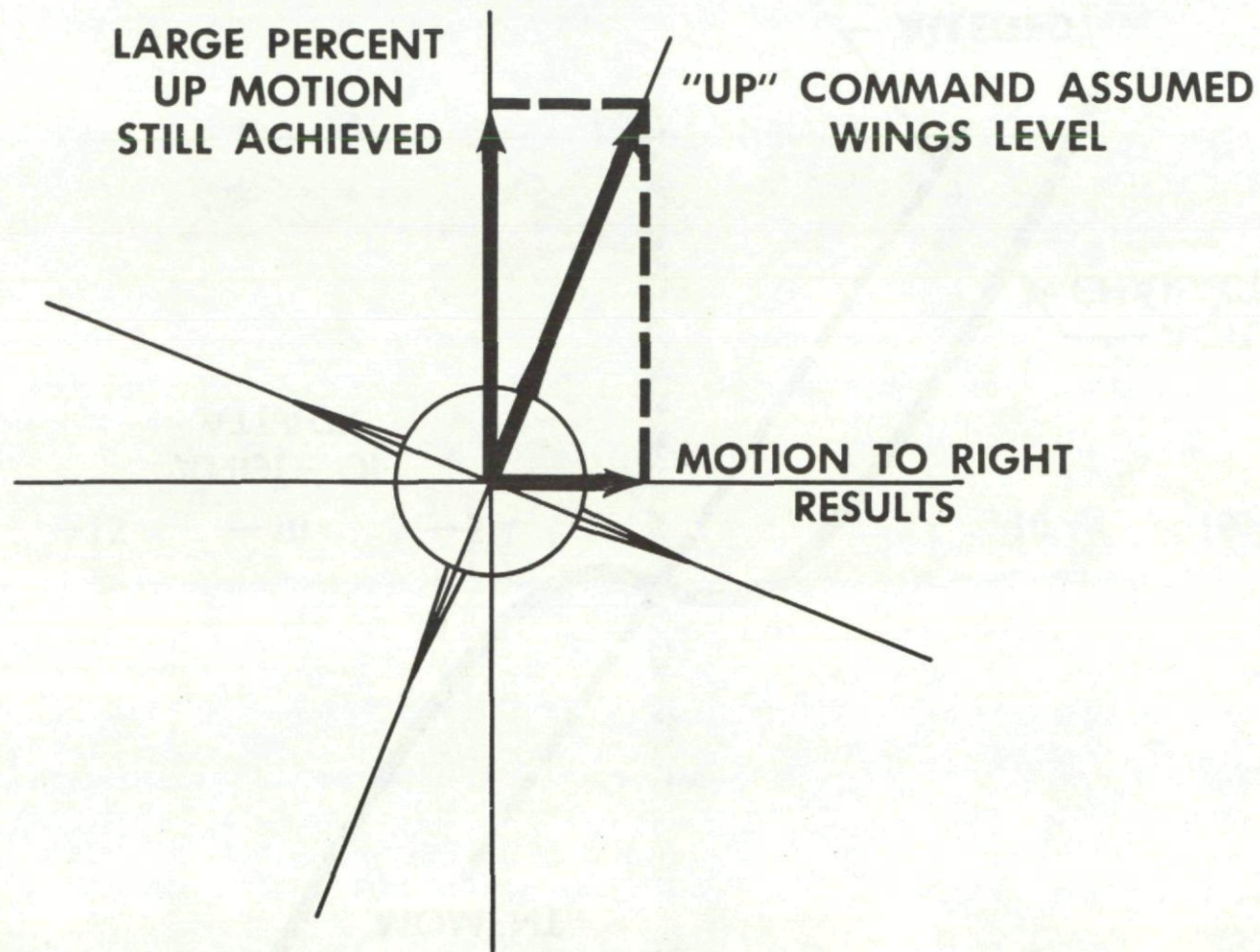


Fig. 6. Cross-talk example.

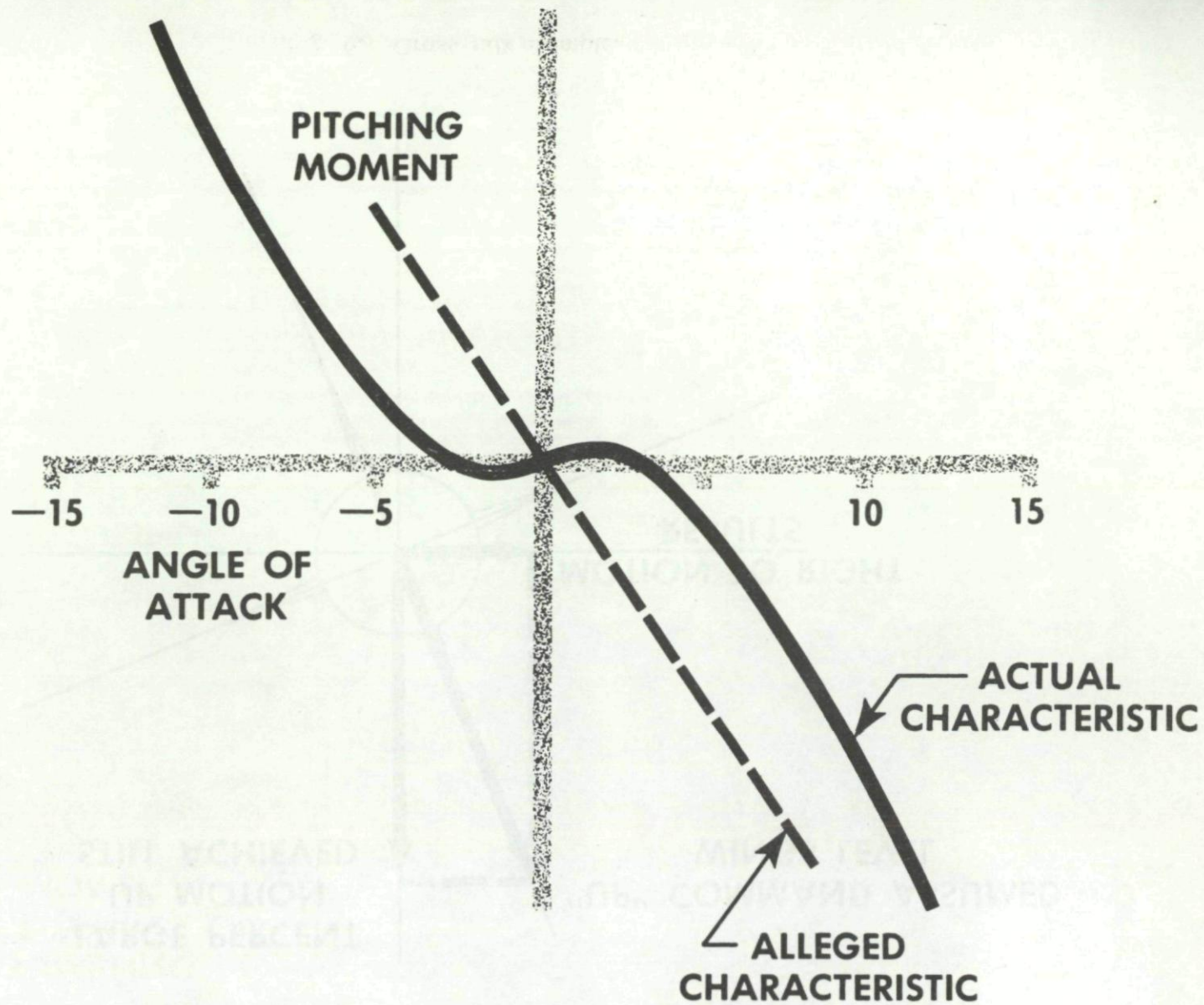


Fig. 7. Pitching moment nonlinearity at small angles of attack.

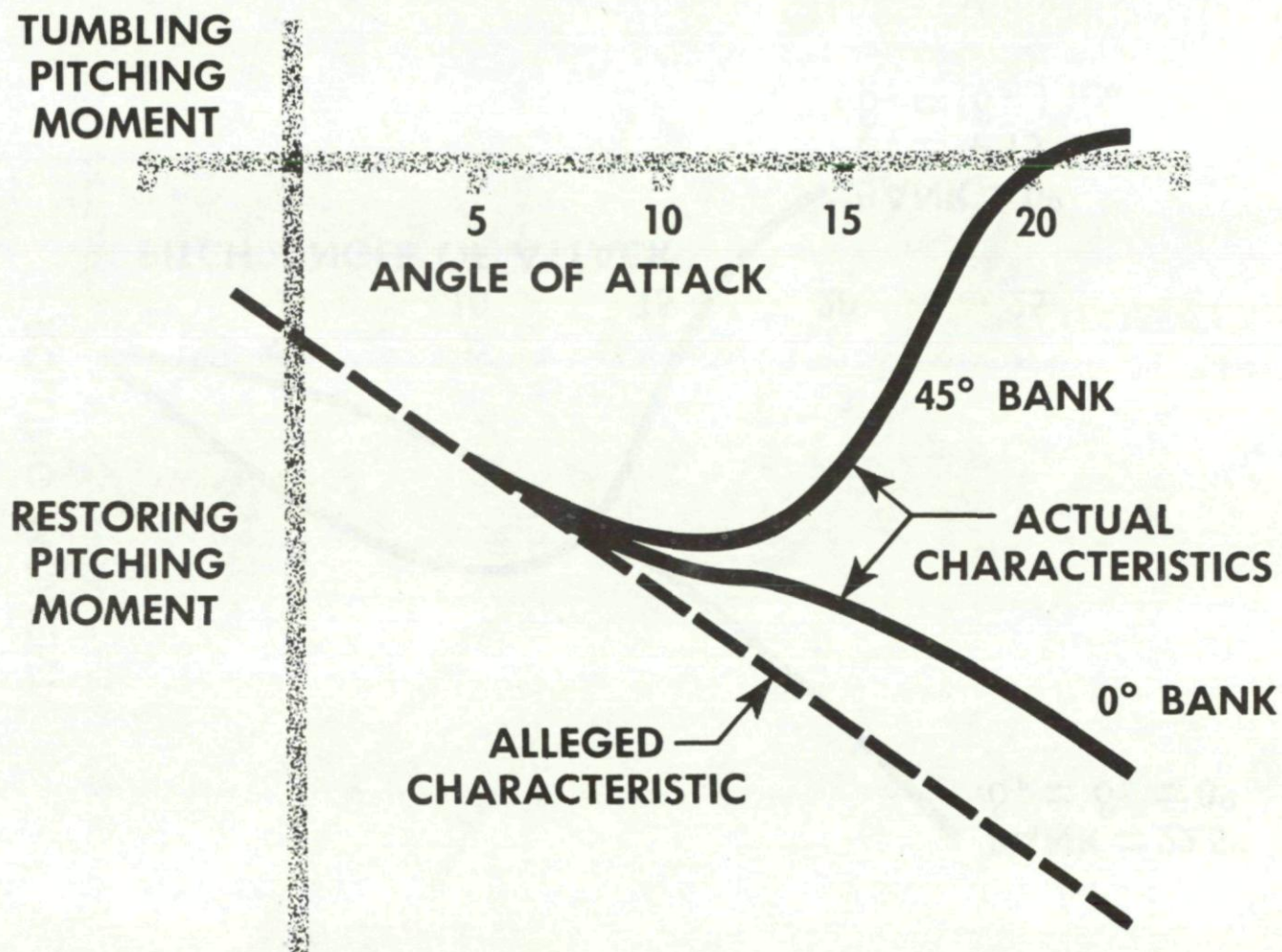


Fig. 8. Pitching moment nonlinearity at bank angle and high angle of attack.

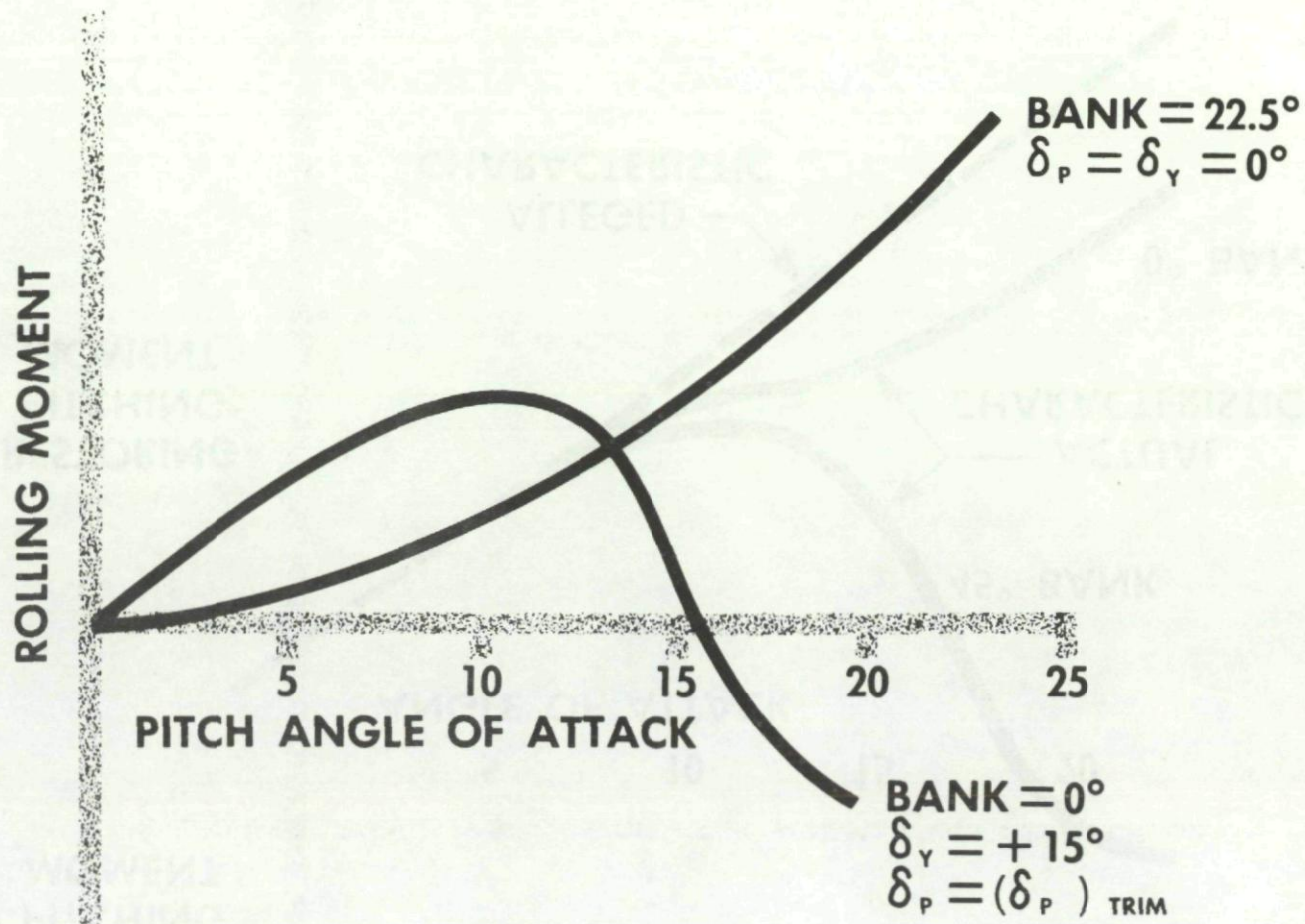


Fig. 9. Rolling moments due to steering.

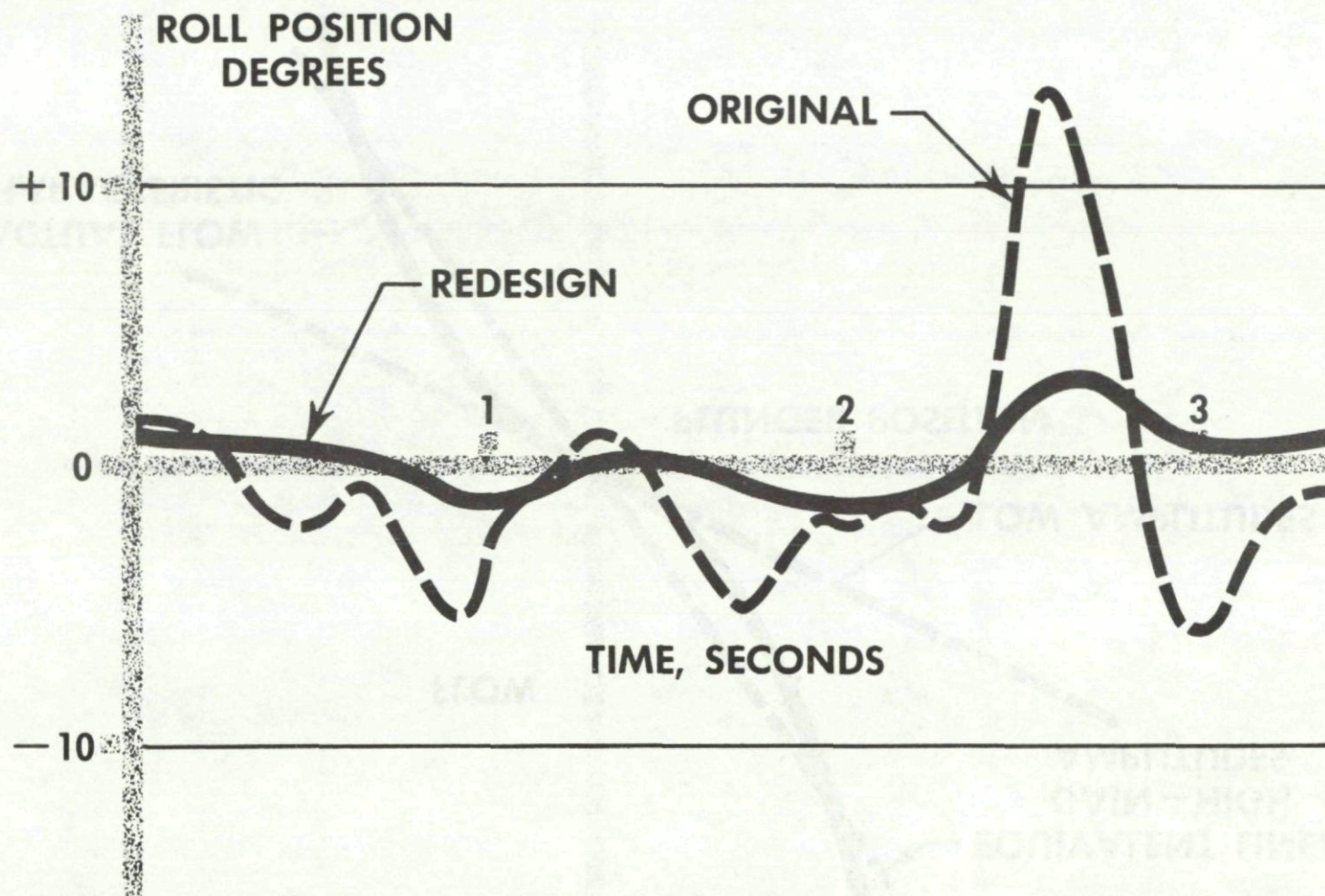


Fig. 10. Induced roll transients.

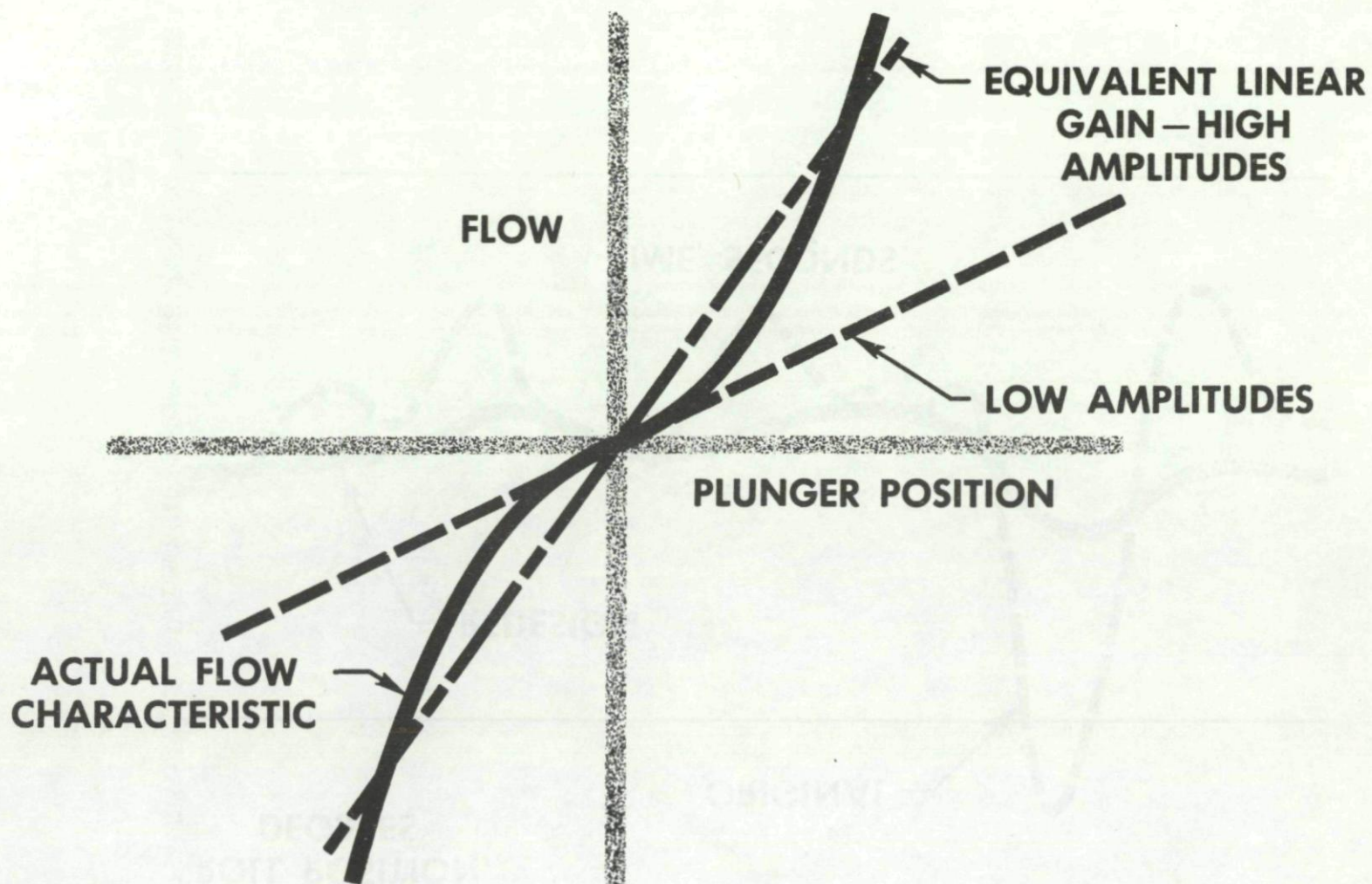


Fig. 11. Hydraulic valve flow nonlinearity.

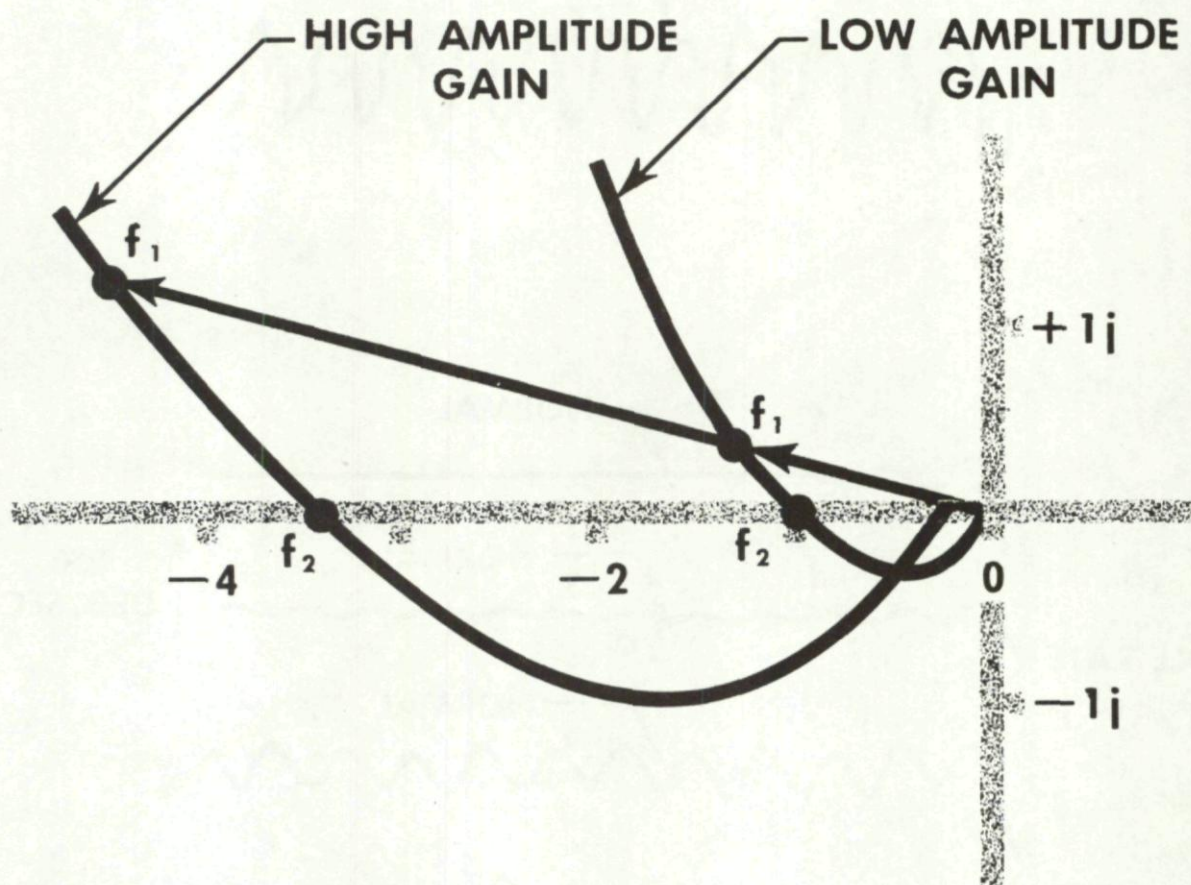


Fig. 12. Nyquist diagram effect of valve nonlinearity.

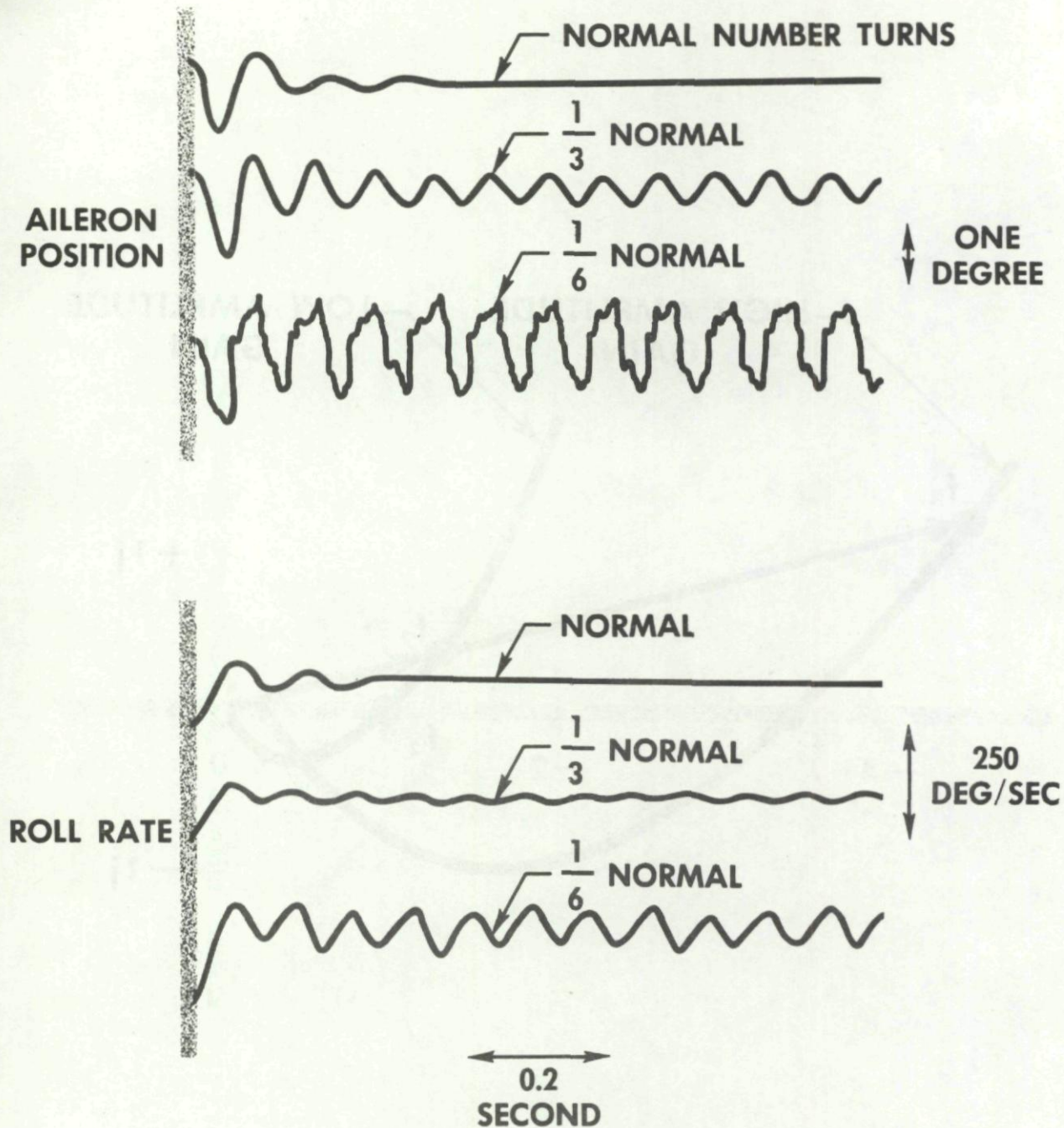


Fig. 13. Effect of aileron potentiometer resolution.

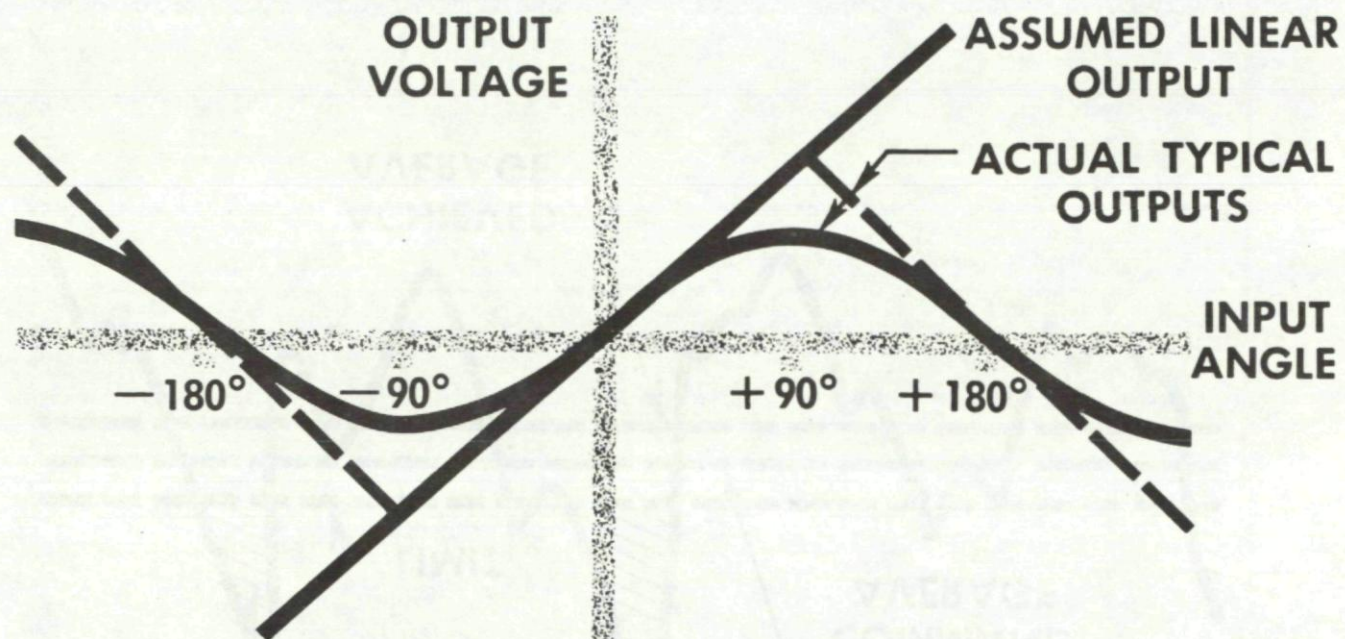


Fig. 14. Typical position pickoff outputs.

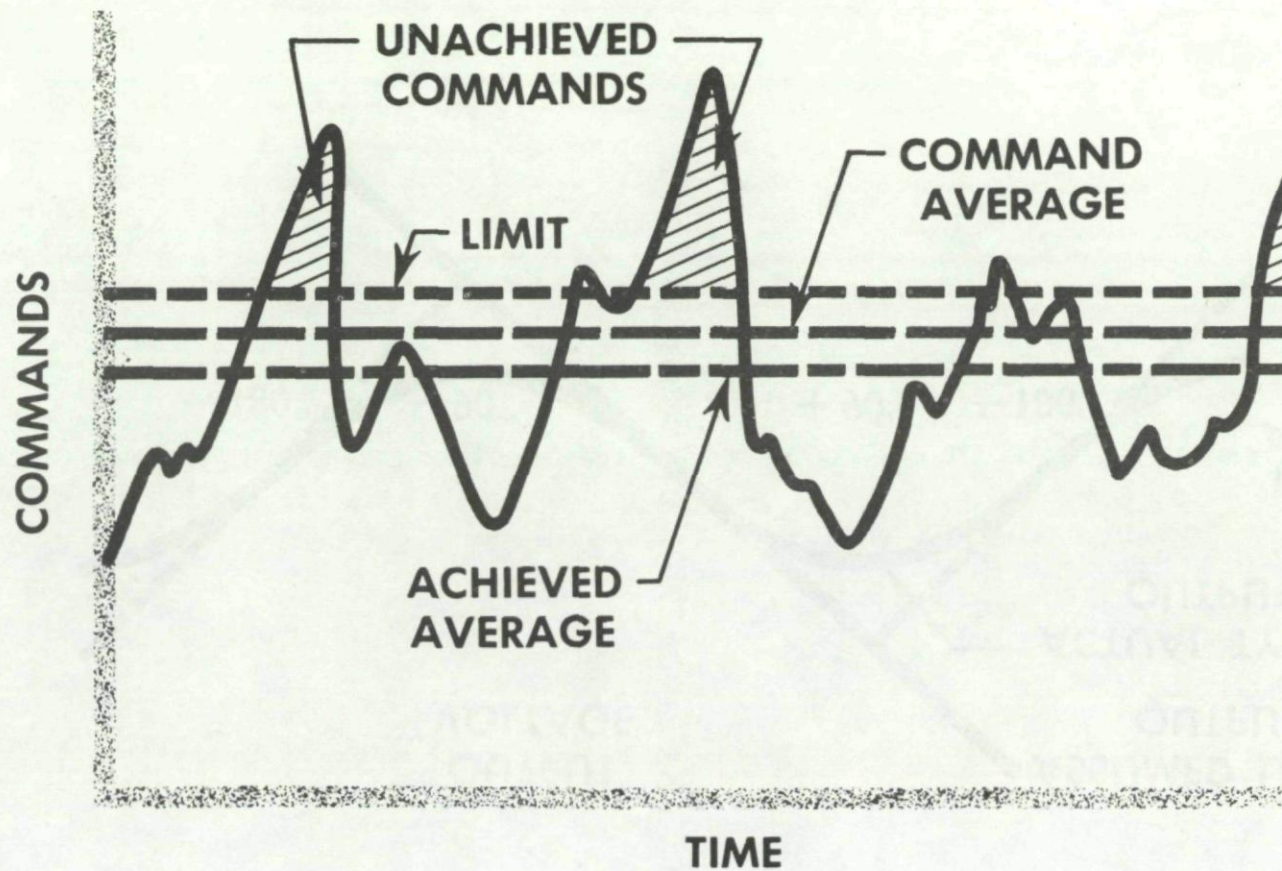


Fig. 15. Effect of limiting on noise.

THE EFFECTS OF AIRFRAME CHARACTERISTICS
ON CONTROL SYSTEM DESIGN
F.E. Perry*

SUMMARY

This paper deals with those aspects of airframe characteristics which affect control system design. Detailed consideration is given to static stability, shift of aerodynamic properties with time-of-flight of the missile, nonlinearities, various types of interaxis coupling, and parasitic feedback due to structural elasticity.

SOMMAIRE

Cette note traite de ces aspects des caractéristiques de structure en vol qui affectent l'étude du système de contrôle. Elle considère en détail la stabilité statique, le changement des propriétés aérodynamiques avec le nombre d'heures de vol du missile, les non-linéarités, les différents types de couplage interaxes, et la contre-réaction parasite due à l'élasticité de la structure.

1. INTRODUCTION

Historically, the piloted airplane has been designed with characteristics which make it easy for the human pilot to control and which make it inherently stable in the absence of an automatic control system. Control systems were first added as a desirable accessory for the prevention of pilot fatigue and, in later years, to improve the accuracy of flight and speed of response under certain tactical conditions. However, even with automatic control systems in use, the desire for good static stability remained, because it was desirable to fly the airplane manually at times, and also because of possible failures of the control system.

In developing airframes for missiles we need to revise our thinking a bit, recognizing certain fundamental differences which exist between the missile and the piloted airplane,

and changing the airframe's characteristics accordingly. Properties which would be quite desirable in a piloted airplane may well make the job of the missile control system designer more difficult.

Perhaps the most important difference between the piloted airplane and the guided missile is that (in most cases) the missile is not required to operate without automatic control. To design a missile which is inherently stable without control would be useless, since a control system failure will generally result in an aborted mission anyway. If this is recognized in the early stages of a missile development program, the overall performance of the missile can be improved, and the job of control system design can be made less difficult. The ability to fly stably without automatic control may prove to be a liability, rather than an asset, for reasons which will be discussed later.

*Boeing Airplane Company, Seattle, Washington.

Another difference is that piloted aircraft are generally designed for positive load factors during maneuvers, while the terminal guidance accuracy of a missile may well be impaired if the same approach is used.

A third major difference between the piloted airplane and the guided missile lies in structural requirements. The performance requirements of the missile may be much more severe than those of the airplane, requiring high gains in the control system. This can lead to difficulty if structural resonances fall within the passbands of control-loop elements.

It is the purpose of this paper to outline briefly some of the airframe characteristics which influence control system design, and to describe their effects. Such a discussion would not be complete, of course, without mention of some of the techniques for handling undesirable characteristics when they exist, because an optimum design from a control system viewpoint will generally not be an optimum design from the viewpoints of the aerodynamics engineer, the propulsion expert, and the structural designer.

Because of these conflicts, it is absolutely essential that the viewpoint of the control system designer be reflected in the preliminary design of a missile, so that a reasonable balance can be obtained between the various factors which govern its overall design. If there is any single idea which should be emphasized, it is that control system considerations should be integrated into the design from the beginning, and not ignored until the airframe design is complete. Failure to do so may result in the addition of innumerable electronic devices, or "black boxes," at some later date, so that the adverse characteristics of the airframe may be overcome. This adds to system complexity, with corresponding decrease in reliability, which is one of the most difficult problems the missile designer must face.

2. AERODYNAMIC CONSIDERATIONS

This section deals with some of the aerodynamic properties of an airframe which influence and limit the design of a missile's automatic control system. The methods used in obtaining desirable control characteristics are properly the domain of the aerodynamicist; hence we will not consider in detail the means of obtaining them. The intent is simply to indicate what the desirable characteristics are so that they may assume their proper place in the preliminary design process in which propulsion, aerodynamic, and control factors must all be considered. Thus the design may evolve as a whole with a realistic balance of factors.

a. Static Stability

Positive static stability has been recognized for some time as an essential quality of the piloted aircraft. Otherwise, even if the rate of divergence were low enough to permit human control, a moment of inattention on the part of the pilot might well result in the aircraft attempting to turn end-for-end, with disastrous results. On the other hand, the missile, with an automatic control system in continuous operation, is not subject to such restrictions. It can be operated with neutral or even negative stability, since the static stability of the airframe is supplanted by the dynamic stability of the overall system. In the process, some beneficial results may arise.

The degree of static stability of an airframe is determined by the distance between its aerodynamic center and its center of gravity (c. g.). Location of the airframe c. g. ahead of the aerodynamic center results in a restoring moment which tends to return the airframe to a neutral position when it is displaced from its trimmed condition, i. e., when a change in angle of attack occurs.

Or stated differently, the slope of the pitching moment vs. lift coefficient curves are such that a nose-down moment exists for nose-up changes in angle of attack from the trimmed condition, and thus a negative slope is required, using the usual notation.

Large amounts of static stability are generally undesirable in an airframe intended solely for automatic control, for the following reasons:

(1) Ease of control and a large degree of static stability are mutually incompatible. Relatively large control-surface angles, which increase the trim drag, will be necessary to trim the airframe to a given lift condition. Also, nonlinearities may be introduced in the pitching moment vs. surface deflection relationship, because of large surface angles. If large surfaces are required, flutter problems may result.

(2) Large hinge moments, and consequently large amounts of installed actuator torque may be necessary to provide the necessary airframe trim angles.

(3) Larger total surface area will be required to provide a given degree of static stability than would be required if the same degree of dynamic stability were provided by a controller; consequently the drag and structural weight can be less if the controller supplies the stability.

(4) In a neutrally-stable, tail-control airframe, the tail may provide useful lift. If a large degree of static stability is present, this is not possible for all trim conditions.

(5) In configurations where it is desirable to use the same surfaces for both pitch (or yaw) and roll control, the large surface trim angles required for some flight conditions could result in insufficient differential surface deflection for roll control.

To see that a controller is capable of providing stability and damping which are interchangeable with the stability and damping inherent in the airframe, let us consider first the so-called fixed-stick airframe equation for angular motion about the pitch axis, and, second, the equation which describes one particular type of controller when installed in the same airframe. Summation of torques about the pitch axis gives, for the airframe only,

$$I_y \frac{d^2\theta}{dt^2} + B \frac{d\theta}{dt} + C\theta = 0 \quad (1)$$

where I_y , B , C , are constants which represent the airframe pitch moment of inertia, its damping coefficient, and its static stability, respectively.

At sufficiently low frequencies we may ignore such things as power-servo, filter, and control-instrument phase lags, and assume the existence of an ideal controller whose response is such that it can instantaneously exert a control moment, D , (due to surface deflection) proportional to the displacement θ , and a moment E , proportional to the body angular rate $d\theta/dt$. We now have,

$$I_y \frac{d^2\theta}{dt^2} + (B + E) \frac{d\theta}{dt} + (C + D)\theta = 0. \quad (2)$$

The damping has been modified by the ratio $(B + E)/B$ and the static stability by the factor $(C + D)/C$. It will be recognized that the factors D and E are subject to control by adjustment of gain constants within the control loops. Thus, even with negative stability in the basic airframe, we can provide positive stability by means of the control

system. Also, in a practical case, the damping provided by the controller may be an order of magnitude or more greater than that which is inherent in the airframe alone. The above case could be extended to include signals proportional to body angular acceleration as well. Apparently, then, the static stability which is designed into conventional piloted airplanes should be applied to missile design only when we wish to fly without automatic control. Neutral stability is a better choice.

b. Center of Gravity Travel and Shifts in Aerodynamic Center

Regardless of the comments above, the dynamics design of a control system can be tailored to any one of a wide range of positive and negative stability values. Trouble arises, however, when large changes in stability occur during flight. These changes may stem from such things as movement of the airframe c. g. as its disposable load is used up, or from shifts in the aerodynamic center as the Mach number varies. A control system design for such an airframe may suffer increased complexity or a compromise in performance for some flight conditions. Therefore, consideration should be given to choosing an airframe configuration which avoids these problems by having the distance between the aerodynamic center and the c.g. remain as nearly constant as possible.

This ideal airframe can only be approached in varying degrees in actual practice. Thus it may be necessary, as velocity and fuel loading change, to program control system constants as a function of time, or to compensate them according to some function of static and total pressures. (The latter may be necessary anyway because of changes in aerodynamic coefficients.) This of course falls in the category of a cure rather than the prevention of the illness itself, but it may avoid difficulty elsewhere in the design. So

we add a compensating device to the system, and system complexity and probability of failure increase. A better choice is to give the problems of c.g. travel and shift in aerodynamic center careful consideration in the preliminary design of the missile.

c. System Nonlinearities

A number of nonlinearities necessarily exist in any airframe and its controller, in the form of such things as the structural limit of the airframe, which cannot be exceeded, and the deflection and velocity limits of control surfaces. In the past several years, methods for deliberately introducing nonlinearities for optimum servo system performance have received considerable attention. As yet well-proven nonlinear design techniques are not generally available, however, and system performance may require compromises wherever nonlinearities appear. Analogs may be used to study the effects of nonlinearities, but the system design usually becomes more cumbersome.

As an airframe is operated at higher and higher altitudes, the problem of developing enough lift to perform maneuvers dictates increasingly large angles of attack. Consequently, unless the aerodynamic characteristics of the airframe are carefully investigated throughout the entire operating envelope of the missile, some regions may be encountered where the control system design job is more difficult. Fig. 1 illustrates one possible type of nonlinear relationship between angle of attack and pitching moment which may exist at high pitch angles of attack. Since the incremental change of pitching moment with angle of attack decreases at the larger lift values, an equivalent change in control-loop gain occurs. Such a change may result in instability unless control-system performance is allowed to deteriorate by using lower gains, or unless

compensating devices are added. A compensating device might take the form of a transducer which senses angle of attack and provides input signals to a multiplier, in the form of a small servo or an electronic circuit. This device would then modify system gain constants as a function of angle of attack. Another "black box" would therefore be added, and system reliability would decrease still further.

Still a third option is apparent, that of limiting the allowable maneuvers in the questionable areas. This, however, is hardly a desirable solution. Steps should be taken, if possible, during the airframe design to insure linear behavior over the entire operating range.

d. Interaxis Coupling

The high velocities and response rates which are typical of high performance missile designs, coupled with operation at large angles of attack at the higher altitudes, has caused airframe behavior to become increasingly critical to interaction between the airframe axes. Thus it becomes more and more desirable that consideration be given during the preliminary design of an airframe to the parameters which cause cross coupling. Neglecting electrical cross coupling, since we are concerned here only with airframe characteristics, the interactions between the various airframe axes are of two types: aerodynamic and inertial. Since both types contribute to unwanted airframe motions, it is appropriate to consider each major cross-coupling term and the available means for eliminating it or reducing its effect.

Techniques for evaluating the effects of cross coupling on the controlled or uncontrolled airframe include the analytical solution of the equations of motion and complete three-dimensional analog simulator studies which include all important cross-coupling terms.

A comparison of the relative values of all the inertial and aerodynamic cross-coupling terms will generally indicate that a number of them may be neglected. Such a study should consider the following terms:

e. Gyroscopic Cross-Coupling

High roll velocities, coupled with long slender bodies for the reduction of drag, result in an increase in the pitching and yawing moments induced by combinations of yawing and rolling velocity, and of pitching and rolling velocity, respectively. The angular accelerations due to this action are:

$$\frac{I_x - I_y}{I_z} p q = \dot{r} \quad (3)$$

and

$$\frac{I_z - I_x}{I_y} p r = \dot{q} \quad (4)$$

where p , q , and r are roll, pitch, and yaw angular velocities respectively; and I_x , I_y , and I_z are the moments of inertia about the missile roll, pitch, and yaw axes.

Since any increase in body length increases I_y and I_z , while smaller body cross sections decrease I_x , the numerators of Eqs. (3) and (4) tend to increase in the supersonic missile design. Thus, the equations of motion should include these terms when analyzing the performance of such missiles. It is probable that other design considerations will permit little flexibility in controlling the size of these terms, but the desirability of keeping roll rates as low as possible consistent with the required missile performance is apparent.

f. Roll Moments Due to Side Velocity

Roll moments due to side velocity may be introduced by wing dihedral, or by unequal lateral drag above and below the longitudinal centerline of the airframe. This characteristic is usually deliberately introduced in a conventional airplane, so that the airplane will automatically bank such that its side velocity is reduced by rolling the velocity vector into the pitch plane. This is appropriate, since load factors are generally applied in the positive direction in the piloted airplane.

The terminal phase accuracy requirements of a missile are often such that it is necessary to develop high values of lift in both the positive and negative directions. It may be shown by solution of the fixed-stick equations of motion for an airframe with roll due to yaw that it will be unstable for some lift conditions and not for others. As a consequence, a high degree of stability in roll for positive loads is not desirable, since this will generally detract from the stability of the airframe under negative load conditions; and the control system must of course be capable of controlling the missile in this latter mode of operation.

Unless we recognize roll due to side velocity as a potential source of trouble in the design of the missile's control system and exercise the necessary care, roll-yaw coupled system instability may result during flight. At best, the complexity of the system may increase as a result of the cross coupling, requiring such measures as the cross-feed of electrical signals between the roll and yaw system electronics. As in the case of negative longitudinal stability, the control system may be able to maintain stable flight over the desired range of negative lift conditions; however, a more straightforward approach is to use a symmetrical airframe when both positive and negative load conditions are to be encountered.

g. Roll Moment Due to Rudder Deflection

In airframes having rudder arrangements which are not symmetrical about the longitudinal axis, rolling moments will be introduced by rudder deflection. Thus, a spurious roll moment is generated by the yaw-axis controller; and this can adversely affect system stability if not taken into consideration in the design of the control system.

Several techniques are available in overcoming the effects of roll-yaw coupling. One of these is to use relatively high gains in the roll system, so that the ailerons react rapidly to counteract roll moments. Unfortunately, the structural stiffness of the wings will generally impose a practical upper gain limit, due to the first asymmetrical bending mode of the wings. Elimination of roll due to rudder deflection can also be accomplished by aileron action in another way. If the aileron servos are properly signaled, through the roll control system, by rudder motion, aileron action may be used to counteract the roll torque generated by the rudder. Mechanical coupling would probably not be desirable, since rudder motion would counteract normal aileron motion unless a unilateral coupling device could be devised. Regardless of the type of coupling used, whether electrical or mechanical, system complexity will increase.

A third, and preferable, solution to the aerodynamic roll-yaw coupling problem is to eliminate it at its source by means of a symmetrical rudder design.

h. Yaw Due to Roll

Rolling an airframe which has an angle of attack, or mushing velocity, tends to rotate its mushing velocity vector into the yaw plane, proportional to the product of

mushing velocity and roll rate. If roll due to yaw exists also, this coupling results in a closed loop which may be unstable for some flight conditions, since roll motion produces further yawing. Reduction of the magnitude of this term can be achieved by operating with the lowest acceptable roll rates, for a given missile design; or by reducing the mushing velocity. Pivoted wings for developing lift could be used to avoid the necessity for large body angles of attack, but with the attendant disadvantage that the body cannot be used to produce lift.

Other cross-coupling terms which may be of interest arise from aileron deflection in the presence of an angle-of-attack, wherein a yawing moment is introduced; or from side loads on an unsymmetrical tail due to roll rate. These terms should be compared numerically with the others present in the system for the expected range of missile motions, and included in the three-dimensional equations if necessary. In the case of yaw due to roll rate, a fin or rudder configuration which is symmetrical above and below the missile longitudinal axis is of course desirable.

3. STRUCTURAL CHARACTERISTICS

In high-performance missiles, relatively high gain control loops are necessary to obtain the speed of response and degree of damping of missile motion which is required. The overall design of a control system for such a missile may begin with separate system designs for the pitch-, roll-, and yaw-axis control systems, using analog equipment and analytical techniques; and then proceed to a complete three-dimensional analog computer simulation of the system, including all the important inertial, aerodynamic, and electrical cross couplings that are present in the system. These studies are based upon the inertial and aerodynamic

parameters of the missile, as obtained by weight calculations, wind tunnel tests, and perhaps preliminary flight tests of some sort. Aerodynamic coefficients may or may not recognize aeroelastic effects, depending upon the designer's assessment of their importance.

The outcome of the above studies is a specification of control-loop transfer functions based upon the behavior of the missile, usually assumed to be a rigid body, throughout the entire velocity and altitude operating envelope of the missile. High performance systems may require both body angular rates and their derivatives as damping signals, in order to obtain the desired performance.

The treatment of the airframe as a rigid body in defining system requirements is a logical and necessary first step. That this departs considerably from the truth will be apparent, however, upon examination of the resonance characteristics of the average structure. In some structures excitation of control instruments by body bending or local structural resonances may introduce parasitic loops which impose more severe restrictions upon system gains than those imposed by the control loops themselves. External appendages, such as engine pods and their attachments, may also form resonant systems which produce the same results.

In order to visualize the nature of this problem, a missile in flight can be considered as a so-called free-free elastic bar or beam. Forces or torques applied to such a beam excite transverse vibrations which, in the simple case of a uniform bar, are given by the relationship (Ref. 1),

$$f = \frac{\pi c k}{8 l^2} (3.0112^2, 5^2, 7^2, 9^2 \dots) \quad (5)$$

where

f = frequency of vibration

ℓ = length of bar (cm)

c = wave velocity in bar material
(cm/sec)

k = radius of gyration of bar (cm)

The higher frequencies are not harmonics of the fundamental, even in the uniform bar; and mode shapes of the first two frequencies are as shown in Fig. 2. Only the odd-numbered frequencies are symmetrical about the center of the bar. Excitation of these modes in the missile body during flight may arise from several sources:*

- (1) Aerodynamic forces or torques on the control surfaces.
- (2) Reaction torques on the structure as a result of control-surface angular accelerations.
- (3) Linear forces applied to the structure as a result of control-surface mass unbalance and angular acceleration.

The generation of these forces which produce body excitation may be visualized by referring to Fig. 3. Actual mode shapes will differ somewhat from those shown in Fig. 2 because of the uneven stiffness and mass distribution of the missile body.

*Sources of excitation, such as engine noise, which do not enter into a closed loop including control electronics, are neglected.

The closed parasitic loop which consists of the missile body, the control instrument, the control system electronics, and the control surface actuator, is shown in Fig. 4, along with the normal control loop. The in-flight gain margin of this system may be predicted on the basis of tests which are carried out in the laboratory. First, a transfer characteristic between the control surface and control instrument, using the inertial force produced by surface acceleration alone, is obtained by electrically driving the actuator. The ratio of total in-flight excitation to that due to inertial effects alone may then be used to predict in-flight behavior.

Excitation of the missile body by the reaction torque of the surface has not proven to be troublesome within the author's experience. However, if reaction torques were troublesome, the influence of torques and linear forces could presumably be separated by temporarily providing artificial mass balance about the control-surface hinge line so that only a torque would be produced by surface acceleration; or by applying a linear force, only, to the fixed surface at the hinge line.

The ratio of total forces which exist during flight to those which are present during ground tests may be readily calculated. This ratio, for small sinusoidal angular surface motions, is

$$\frac{F_m + F_A}{F_m} = \frac{m \ell \omega^2 + Z \delta}{m \ell \omega^2} \quad (6)$$

where

F_A = aerodynamic force due to surface deflection

F_m = inertial force due to surface angular acceleration

m = equivalent mass at surface c.g.

ℓ = distance of surface c.g. from hinge line

Z_δ = the linear force coefficient per radian of surface deflection

ω = 2π (frequency of surface excitation)

The in-flight behavior of the system can therefore be predicted from the results of a simple test which is carried out in the laboratory, plus known aerodynamic coefficients.

Since the parasitic loop gains are high only for high-Q structural resonances, the frequencies of concern are only those which correspond to the natural bending frequencies of the body.

An experimental technique for obtaining gain-margin values which immediately appears attractive consists of increasing the gain of the control loop electronics, under laboratory conditions, until instability occurs, and then modifying the gain margin figure thus obtained by the ratio given by Eq. (6). This may or may not give the correct results, since the inertial forces increase (for a given amplitude) as the square of the surface excitation frequency, while aerodynamic forces remain almost constant at low frequencies. Consequently, the frequency at which instability occurs due to the inertial force alone is not necessarily that at which instability would occur during flight, since the structure will have several resonant modes of vibration.

Recognizing that body bending may be a problem, several means are at our disposal for avoiding trouble:

- (1) We may design the missile structure such that its resonant frequencies are above the passband of the control-loop elements.
- (2) We may locate the control instruments, if possible, at nodal points.
- (3) We may use special methods for producing control signals which represent rigid body motions, but which are not direct measurements.

With regard to (1) above, weight limitations may preclude a structural design with resonant frequencies outside the passband of the elements in the servo system. Filtering could be used to deliberately decrease passbands consistent with the allowable control-loop phase shifts, but usually only with some degradation in system performance. Therefore a "quiet" location for the control instruments becomes attractive.

Reference again to Fig. 2 indicates a possible location at point A where body attitude, body angular rate, and body angular acceleration measurements which are free from body vibration at the fundamental body-bending frequency may be taken. Similarly, linear acceleration transducers could be located at point B. Unfortunately, no common nodal point exists for both the first and second bending frequencies in either case. (This may be possible in the case of the actual missile structure, which is more complex.)

One possible approach is that of combining (1) and (2) above. The selection of optimum transducer locations may be sufficient at frequency f_1 ; and excitation at the second body resonant frequency can very likely be attenuated by filters, since it will probably lie above the passbands of the control elements. Attention has not been given to the higher frequency modes, since the attenuation characteristics of control instruments, actuators, and other components can be expected to prevent difficulty at these frequencies.

A second approach which has been used successfully in overcoming parasitic-loop instability is that of obtaining "computed" body angular-rate and angular-acceleration signals from control surface deflections over a limited frequency band. For instance, the angular acceleration of a missile about its pitch axis is given by the expression:

$$\ddot{\theta} = M_q q + M_w w + M_{\delta} \delta_E \quad (7)$$

where

q = pitch angular rate

δ_E = elevator deflection

w = mushing velocity

and M_q , M_{δ} , and M_w are the angular acceleration coefficients due to missile pitching rate, elevator deflection, and missile mushing velocity, respectively, for the particular airframe in question.

The presence of the term $M_{\delta} \cdot \delta_E$ provides a means of computing $\ddot{\theta}$, to a close approximation, if the other terms are sufficiently small. (The coefficient M_w will be recognized as depending upon the distance between

the aerodynamic center and the airframe c.g., and would be small for near neutrally stable missiles.) In a similar fashion, integration of surface deflection would permit an approximation of body angular rate. The method described here would not, of course, be suitable at zero frequency where steady-state surface trim angles are required.

Several methods have been outlined above for overcoming difficulties associated with structural resonances. These resonances, if low enough in frequency, may define the upper limit of control-loop gains due to the presence of parasitic loops. Methods for overcoming this difficulty usually lead to increased control system complexity; therefore structural rigidity which is compatible with the performance requirements of the missile should be designed into the airframe if possible.

4. CONCLUSIONS

In order to avoid adverse control characteristics which may later complicate control system design, careful consideration should be given to those missile parameters which affect control in the preliminary design phase of a missile program. This is particularly desirable in the case of high-performance interceptors, where accuracy and speed of response are of primary importance.

Static stability requirements should be reviewed objectively and should not be unduly influenced by conventional airplane practice unless it is desired to operate the airframe for periods of time without a controller. Attention should also be given to minimizing changes in static stability which may result from shifts in the airframe aerodynamic center with Mach number and shifts in c.g. with changes in fuel loading or other forms of weight disposal.

Also of major concern are such items as interaxis couplings and nonlinear behavior of aerodynamic coefficients. These may lead to system instability at some flight conditions unless system design and analysis is very complete; and at best will complicate design or compromise performance.

Structural characteristics should be compatible with the gain levels which are required to obtain the desired control system performance.

Compensating schemes can usually be found for overcoming the effects of adverse

control characteristics on system design. Unfortunately, these lead generally in the direction of increased system complexity with its attendant decrease in system reliability.

5. ACKNOWLEDGEMENT

The efforts of members of the Boeing Airplane Company Applied Physics Staff, Aerodynamics Staff, and XIM-99 Guidance Project, in contributing ideas and reviewing the material for this presentation, are gratefully acknowledged.

REFERENCES

1. Kinsler, L. E., and Frey, A. R., "Fundamentals of Acoustics," Wiley and Sons, 1950.

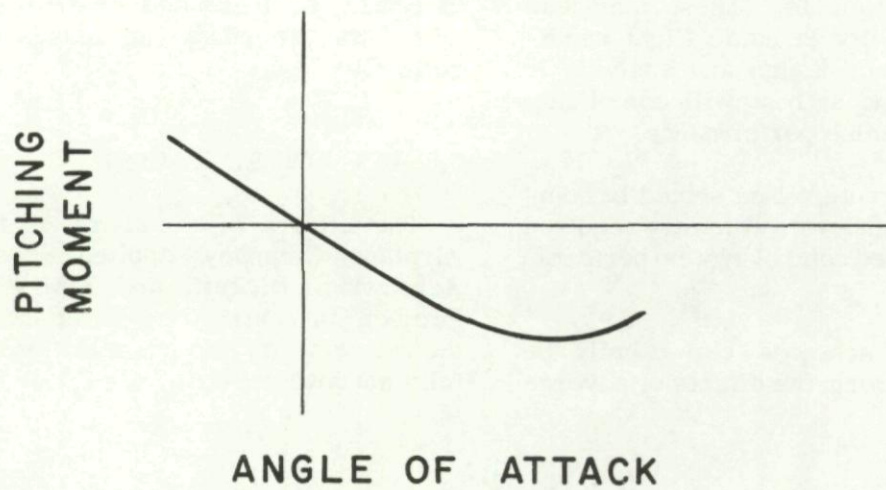


Fig. 1. Undesirable nonlinear relationship between pitching moment and angle of attack.

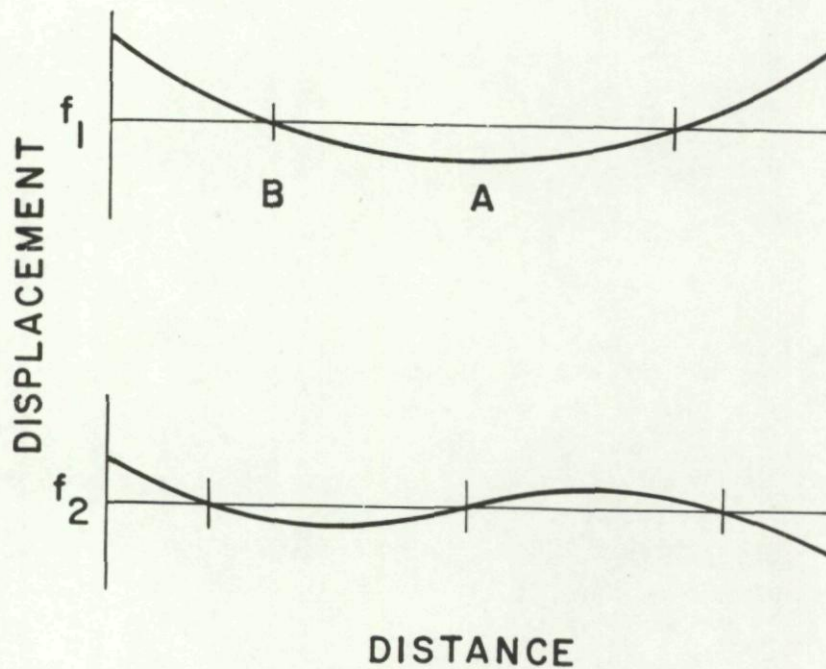


Fig. 2. Transverse bending in a uniform elastic bar.

$T_{A \pm J}$ = AERODYNAMIC PLUS INERTIAL TORQUE

F_A = SURFACE LIFT FORCE

$\dot{\delta}$ = SURFACE ANGULAR ACCELERATION

δ = SURFACE DEFLECTION

F = FORCE DUE TO LINEAR ACCELERATION OF EQUIVALENT MASS m .

l = DISTANCE OF SURFACE c.g. FROM HINGE LINE

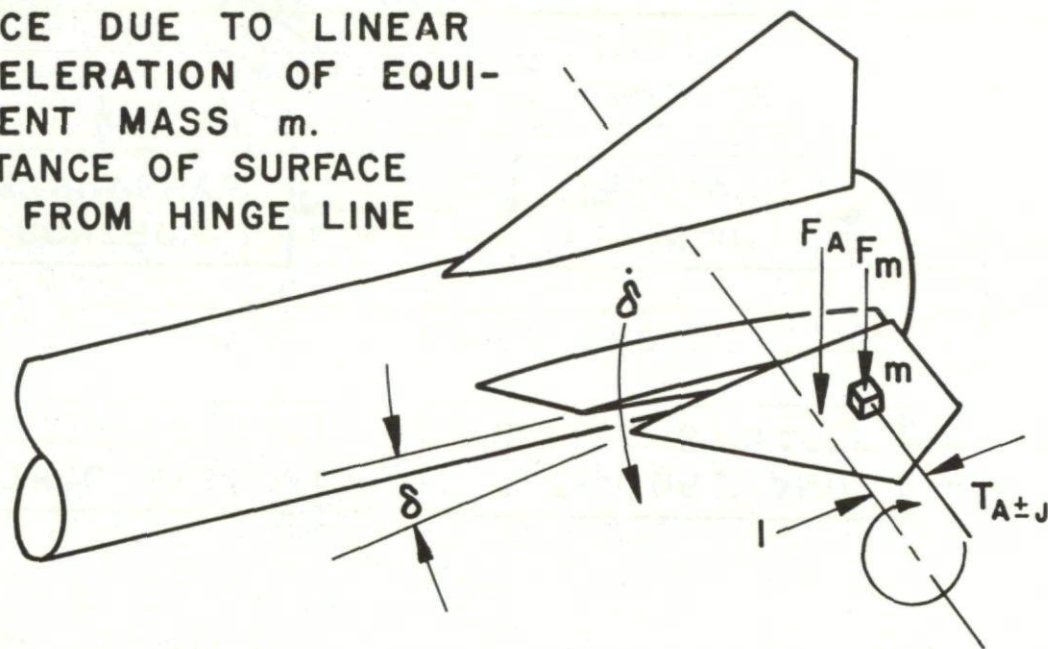


Fig. 3. Torques and forces due to control surface angular acceleration and deflection.

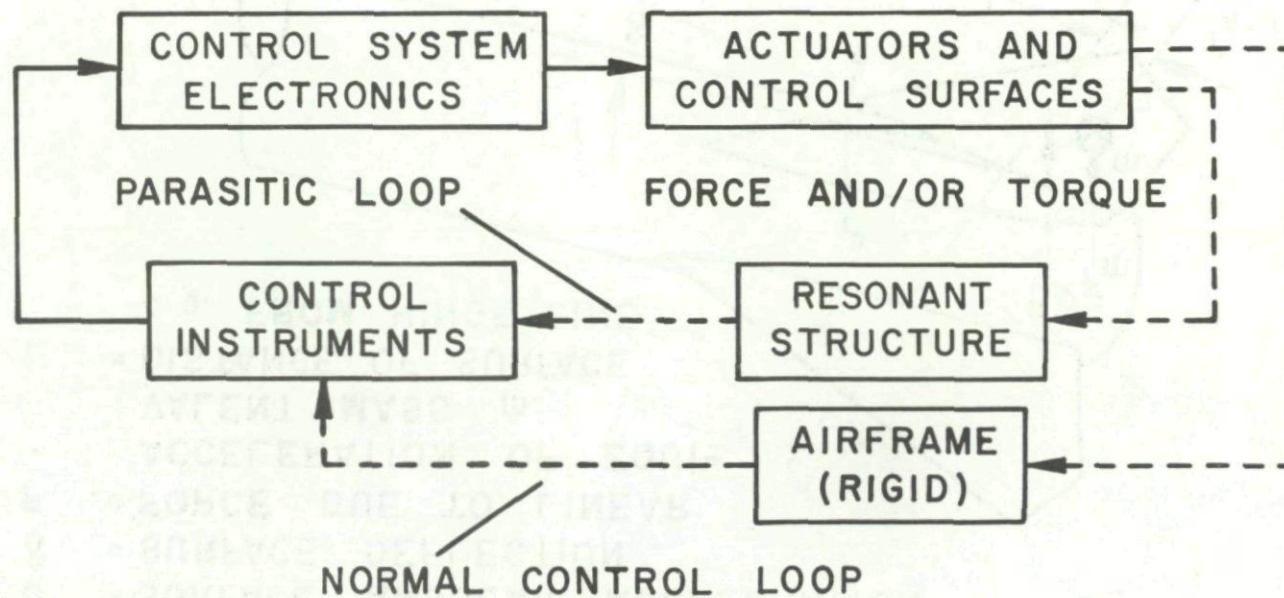


Fig. 4. Parasitic loop due to body bending.

GEOMETRICAL STABILIZATION BASED ON SERVODRIVEN GIMBALS AND INTEGRATING GYRO UNITS

Charles S. Draper and Roger B. Woodbury*

SUMMARY

The paper describes the general problem that must be solved by geometrical stabilization equipment to be carried by land, sea, and air vehicles. A solution for the stabilization problem by means of single-degree-of-freedom integrating gyro units carried by a three-degree-of-freedom servodriven gimbal system is described by geometrical and functional diagrams. A comprehensive system of concepts, terminology, and symbols adapted for setting up a complete mathematical formulation of the performance equations for a three-gimbal-axis stabilization system is applied to an illustrative stabilization system. The design features and performance characteristics of single-degree-of-freedom integrating gyro units suitable for use in systems of this kind are reviewed. From information summarized in the functional diagram for a typical gimballed stabilization system, the performance equation for motion about a single gimbal axis is set up in a form that may be conveniently adapted for the complete system when drives for all three axes are in simultaneous operation. To illustrate one application of the performance equation, it is shown that the use of gear-train drives for the gimbals is not as favorable for high-quality stabilization performance as the use of direct-drive motors.

SOMMAIRE

Cette note décrit le problème général qui doit être résolu par l'équipement de stabilisation géométrique, devant être transporté par route, mer et air. Une solution au problème de stabilisation par le moyen d'appareils gyroscopiques intégrant à un seul degré de liberté porté par un dispositif asservi de mise à direction fixe à trois degrés de liberté est décrite à l'aide de diagrammes géométriques et fonctionnels. Un ensemble compréhensif de concepts, de terminologie et des symboles adaptés pour une mise en équations mathématiques complète des équations de performance pour un système de stabilisation par un dispositif de mise à direction fixe par trois axes, est appliqué pour illustrer un système de stabilisation. Les caractéristiques étudiées et les caractéristiques de performance d'appareils gyroscopiques intégrant à un degré de liberté convenables pour l'utilisation dans des systèmes de ce genre, sont revues. A partir de l'information résumée dans le diagramme fonctionnel d'un système de stabilisation typique par dispositif de mise en direction fixe, l'équation de performance pour le déplacement suivant un seul axe de mise en direction fixe est écrite dans une forme qui peut être adapté d'une manière intéressante pour le système complet, quand des déplacements suivant les trois axes s'opèrent simultanément. Pour illustrer une application de l'équation de performance, il est montré que l'utilisation de train d'engrenages pour l'entraînement des mises en direction fixe n'est pas aussi favorable pour une excellente performance de stabilisation que l'utilisation de moteurs entraînant directement.

*Massachusetts Institute of Technology, Cambridge, Massachusetts.

1. INTRODUCTION

Accurate information on geometrical orientation is necessary for the guidance and control of modern vehicles. Radiation links extending to objects outside moving vehicles are useful, but do not, in general, supply precise and continuous geometrical reference information. Data of this kind may be based on an inertial space, without the need of external physical connections, by the application of gyroscopic principles. This is possible because Newton's laws of dynamics are true with respect to inertial space, so that a spinning rotor tends to hold its plane of rotation in this space and may be used as an orientational reference for any noninertial coordinate system that exists in a moving vehicle. Examples that illustrate this use of gyroscopic action are common in present-day flight instruments for aircraft and navigation equipment for marine vessels. Aircraft that are licensed for all-weather flying must have gyro turn indicators in addition to climb-and-bank indicators, and all except small ships carry gyrocompasses. The purpose of these self-contained gyroscopic devices is to establish orientational reference coordinate systems that are continuously available for indicating or automatically controlling orientation of the craft in which the reference is carried.

Until recent improvements in air- and sea-vehicle performance introduced more stringent requirements on operation, gyroscopic devices depended on the principle of overpowering interference effects in gimbal suspensions by the use of heavy rotors spinning at high speed. Recent developments have shown that it is possible to achieve geometrical stabilization of improved quality from small gyroscopic units that generate electrical signals for servodrive motors of sufficient power to overcome interference effects in gimbal bearings. In practice, this

gyroscope servodriven gimbal combination must provide two functions:

- (1) Stabilization, that is, holding a geometrical reference free from rotation with respect to inertial space in the presence of disturbing torques and arbitrary movements of the support member that carries the stabilization equipment.
- (2) Response to orientational commands, that is, changing the orientation of the reference member with respect to inertial space in response to command inputs.

The aircraft turn indicator is an example of a device that provides both these functions. This instrument stabilizes the indication of north or any selected azimuth direction against oscillations of the airplane in which it is carried. It also changes the reference direction in response to commands from pilot director equipment or to manual adjustments of a setting knob.

It is the purpose of this paper to discuss the mechanical features and operating characteristics of orientational reference systems based on single-degree-of-freedom integrating gyro units supported by servodriven gimbals and designed to be capable of changing the orientation of the indicated geometrical reference in response to command inputs. In order to make it possible for the complex geometrical, instrumental, and dynamical situations involved to be effectively understood and analyzed, self-defining notation and functional concepts are used in setting up the physical and theoretical aspects of the problem (Ref. 1). As a means of keeping this paper to a reasonable length, the phases of operation that are associated

with responses to command inputs will not be considered, leaving the available space free for developing the generalized theory of one type of gimbaled stabilization system and a limited discussion of certain phases of its performance.

2. THE PROBLEM OF GEOMETRICAL STABILIZATION

Fig. 1 shows the essential concepts associated with geometrical stabilization without regard to the mechanical details by which working equipment may be achieved in practice. The support member is the part of a vehicle that carries the stabilization equipment. For convenience in developing geometrical relationships, a set of right-handed Cartesian coordinates, $X_{(sm)}$, $Y_{(sm)}$, $Z_{(sm)}$, is associated with the support member.

In Fig. 1, the reference axes, $X_{(ref)}$, $Y_{(ref)}$, $Z_{(ref)}$, represent the reference orientation of a set of axes, $X_{(cm)}$, $Y_{(cm)}$, $Z_{(cm)}$, fixed to the controlled member that would exist in a perfectly operating stabilization system. The reference orientation that is available in any actual system is given by the indicated reference axes, $X_{(ref)(ind)}$, $Y_{(ref)(ind)}$, $Z_{(ref)(ind)}$, which are identical with the axes fixed to the controlled member. In operation, the desired result of stabilization system operation is to keep the controlled member axes accurately aligned with the reference axes.

3. GIMBAL SYSTEM FOR CONTROLLED MEMBER SUPPORT

The diagram of Fig. 2 shows a three-gimbal system for supporting a controlled member having three degrees of angular freedom with respect to the base which

carries the gimbals and which, in turn, is mounted on a support member. The general arrangement is that described in United States patents issued to the authors of this paper (Refs. 2 and 3).

The controlled member is fixed to the inner gimbal. This member has rotational freedom with respect to the inner gimbal about an axis of symmetry, $Z_{(cm)}$, which is the direction chosen for the z-axis of the inner gimbal, $Z_{(ig)}$. The inner gimbal is carried by the middle gimbal with rotational freedom about an axis perpendicular to $Z_{(ig)}$. The z-axis of the middle gimbal, $Z_{(mg)}$, is taken as identical with the z-axis of the inner gimbal. The y-axis of the middle gimbal, $Y_{(mg)}$, is taken along the axis of rotation between the middle gimbal and the inner gimbal. The x-axis of the middle gimbal, $X_{(mg)}$, is chosen so that it completes a right-handed set with $Y_{(mg)}$ and $Z_{(mg)}$ and so that it is parallel to $X_{(cm)}$ when $Y_{(cm)}$ is parallel to $Y_{(mg)}$.

The middle gimbal is carried by an outer gimbal with rotational freedom about a direction at right angles to $Y_{(mg)}$. This direction is chosen as $X_{(og)}$, the x-axis of the outer gimbal. The y-axis of the outer gimbal, $Y_{(og)}$, is taken along $Y_{(mg)}$; and $Z_{(og)}$ forms the third member of a right-handed-axis set with $X_{(og)}$ and $Y_{(og)}$. The axes of the base which carries the gimbal system, $X_{(b)}$, $Y_{(b)}$, $Z_{(b)}$, are identical with the common directions of the gimbal and controlled member axes when all the gimbals are rotated to positions in which all the x-axes, y-axes, and z-axes are aligned.

4. SINGLE-DEGREE-OF-FREEDOM INTEGRATING GYRO UNITS

Stabilization of a controlled member with respect to inertial space requires some means for rapidly and accurately detecting

deviations of the controlled member with respect to some desired inertial space reference orientation. Because inertial space does not have any natural references in position except the fixed stars, which are not useful without line-of-sight information, any reference orientation for the controlled member of a stabilization system must be established by elements within the system itself. This means that gyro units for stabilization purposes must generate signals that represent angular deviations from a reference orientation established by operation of the gyro rotors. Single-degree-of-freedom gyro units receive such deviations as rotations about a single direction fixed to the units. To generate the signals required to produce complete stabilization, three single-degree-of-freedom gyro units are required for a three-axis gimbal system like that of Fig. 2. The general features of gyro units suitable for use in systems of this kind are briefly described below.

Single-degree-of-freedom integrating gyro units are a development of the Instrumentation Laboratory at the Massachusetts Institute of Technology. These units have been discussed in publications (Refs. 4 and 5) which give bibliographies of sources of theoretical and practical information on gyroscopic theory and instruments using gyro principles, and are also described in United States patents (Refs. 6 and 7). A brief review of the features and essential performance characteristics of the single-degree-of-freedom integrating gyro unit is included here to show how units of this type operate as components of stabilization systems.

Fig. 3 is an illustrative pictorial diagram showing the features of the M. I. T. integrating gyro units. A gyro rotor spun at constant speed by an alternating-current synchronous motor is mounted in a gimbal enclosed by a hermetically sealed cylindrical shell filled with helium to act as a neutral

atmosphere and to serve as a medium for producing an even temperature distribution over the parts of the shell. This floated gimbal has its weight adjusted so that it is almost perfectly supported by buoyant forces in a fluid that completely fills the clearance volume between the float and the hermetically sealed case of the unit. The float is pivoted within the case by watch-jewel-type bearings that, because of the low residual loads, introduce substantially no friction to resist float rotation.

The pivots are carried on axial extensions from the float. On one end, the extension carries the rotor of an alternating-current signal generator of the microsyn (Ref. 8) type that produces signals of the phase reversal type having magnitudes proportional to the angular deviation of the float from the position in which the output voltage has its null level. The end of the float away from the signal generator carries the rotor of a microsyn torque generator which receives electrical current inputs and applies a corresponding torque to the float about the gimbal axis. Balance nuts adjustable from outside the case are used to put the float into accurate rotational balance.

In any actual single-axis gyro unit, a number of features must be incorporated that, in the interests of simplicity, are omitted from Fig. 3. For example, the flexible leads used to make electrical connections from the external terminals to the gyro drive motor are not shown. It is important that the torque applied by these rotor-drive power connections to the gimbal be kept within tolerable levels.

The required performance is achieved by using thin, flat wires of good mechanical properties formed into semicircles which, in their undistorted shape, accurately touch the

terminals on the case and on the gimbal float. In practice, lead-in systems with these features impose satisfactorily low torques on the gimbal. Neglecting the acceleration torque effects that act during transient periods, an angular velocity of the unit case about the input axis, which is at right angles to the gimbal axis and the spin axis when the signal output is at its null level, causes the gyroscopic rotor output torque to act on the gimbal.

This rotor output torque is the result of the tendency of the rotor to turn toward the input axis through the smallest angle in the direction that would finally cause the rotor spin velocity to have the same direction as the input axis angular velocity if the gimbal were allowed to turn into complete alignment with the input axis. Static equilibrium is reached when the resisting torque applied to the gimbal by the integrating damper due to the fluid in the clearance between the float and the case just balances the gyroscopic output torque. This situation occurs when the rate of rotation of the gimbal with respect to the case is sufficient to generate a resisting torque due to viscous shearing action that is equal in magnitude to the magnitude of the torque from the gyroscopic element.

The rate of change of the output voltage from the signal generator is directly proportional to the angular velocity of the gimbal with respect to the case. The net result of the action of all the components except the torque generator is to produce a rate of change of the gyro unit output signal that is proportional to the rate of change of the case orientation with respect to inertial space about the input axis. Integration of the input angular velocity and the output voltage rate shows that the change in the output signal for any time interval within which the

unit is operating properly* is proportional to the angular displacement of the case about the input axis that occurs during the same time interval.

In making practical applications, it is convenient to associate certain definitions, conventions, and symbols with single-degree-of-freedom integrating gyro units, which are also called single-axis integrating gyro units. These definitions and symbols are summarized in Fig. 4.

One form of the performance equation for the integrating gyro unit given in Ref. 4 appears as Eq. (1) of Information Summary 1 in terms of the angular velocity input - voltage output performance function for the gyro unit. In this equation, the performance function has a role that is identical with the transfer function (Ref. 9) of Laplace transform theory. Of the several inputs to which the gyro unit responds, Eq. (1) includes the effects of only three. The actuating inputs for the unit are the angular velocity of the case with respect to inertial space about the input axis, which acts through the gyro rotor, and the input current, which serves to apply orientational control commands to the unit through the torque generator.

Angular acceleration of the case with respect to inertial space about the output axis is an interfering input that produces an undesirable output signal component. This signal component exists because the inertia reaction of the gimbal causes it to lag behind the case rotation against the accelerating torque of the viscous shear coupling between the case and the float.

*When the gimbal is in contact with stops set in the case, the damper is effectively locked, and the unit is not operating properly.

For the purposes of stabilization, the desired function of a gyro unit is to produce an output signal that depends only on angular motion of the case about the input axis. The gyro unit performance equation is restricted to this situation by omitting the input-current term from Eq. (7) of Information Summary 1. The absence of other terms in this equation which would represent various drift-producing torque components that exist in practice means that such effects are assumed to be negligible from the standpoint of stabilization performance.

Eqs. (5) through (14) of Information Summary 1 outline the steps required to change the gyro performance equation to a form adapted for describing the behavior of the gyro unit under steady-state sinusoidal variations of input and output quantities. In a form that is often useful, the performance function depends on the characteristic time, forcing frequency product as the essential dimensionless parameter. For convenience, the performance function is broken down into the angle-voltage reference sensitivity, the angle-voltage dimensionless sensitivity ratio, and the angle-voltage dynamic response angle.

The definitions and symbols for these quantities are given in complete form for the integrating gyro unit because they illustrate a pattern that is generally used for describing the performance of all components that are required to form a geometrical stabilization system.

The second term within the brackets of Eq. (9) of Information Summary 1 shows the effect of angular displacement about the output axis as compared to angular displacement about the input axis. The operator, p , indicates that the effect is 90 degrees out of phase with input axis displacements. The dimensionless forcing frequency - angular acceleration - angular velocity reference

sensitivity ratio product is a measure of the magnitude of the output axis angular acceleration effect. In practice, the sensitivity ratio is small, so that output axis displacement effects will ordinarily not be serious except at high forcing frequencies.

5. ILLUSTRATIVE SINGLE-AXIS SERVO-DRIVEN CONTROLLED MEMBER WITH INTEGRATING GYRO

To illustrate the basic action of an integrating gyroscope servodrive combination for providing stabilization, Fig. 5 is drawn as a line schematic representing the essential features of a controlled member carrying a single integrating gyro unit with its input axis along the axis of rotation of the servodrive system. With no input current applied to the gyro unit and the gimbal in the position at which the gyro output signal has its null level, any rotation of the controlled member about the input axis will cause the gimbal to turn with respect to the case about the output axis. The resulting angular displacement is picked up by the signal generator and sent through sliprings to the power control system.

The output of the power control system acts on the drive motor to produce a torque on the controlled member to cause rotation in the proper direction to force the gyro gimbal back toward its null-output voltage position. The continual repetition of this servo action causes the controlled member to remain close to a position that is the inertial space reference orientation for the gyro unit case.

The reference orientation of the controlled member may be changed by introducing a command signal in the form of an input current to the gyro unit. This current causes the torque generator to apply a torque that starts to turn the gimbal away from its

null signal position and produces a corresponding gyro unit output signal. This signal is the input for the servo system and causes the drive motor to turn the controlled member in the proper direction for the gyro rotor to generate an output torque acting against the output from the torque generator.

For a constant input current, the steady-state condition is reached when the servo drives the controlled member at a rate for which the gyro rotor torque output just balances the torque generator output torque. By making use of this action, the controlled member reference orientation may be changed at will. In any given situation, when the desired orientation is reached, the input current is reduced to zero; the gyro unit servo combination will then stabilize the controlled member with respect to the new inertial reference orientation.

6. GYROSCOPE AND SERVO DRIVE COMBINATION FOR STABILIZATION OF A GIMBAL-SUPPORTED CONTROLLED MEMBER WITH THREE DEGREES OF FREEDOM

Geometrical stabilization of a controlled member with three degrees of freedom may be achieved with three single-degree-of-freedom gyro units mounted on a controlled member with their input axes along three mutually perpendicular axes. The servo action about each axis is generally similar to that described in the last section except for the fact that, with three degrees of freedom, a number of geometrical interactions among the three component gyro unit, servo-drive combinations must be taken into account.

Fig. 6 illustrates the result of a systematic procedure for orienting three gyro units with respect to the gimbal-supported controlled member shown in the diagram of

Fig. 2. The X-gyro unit is fixed with its input axis pointed along the direction of $X_{(cm)}$, its output axis along $Y_{(cm)}$, and its spin reference axis along $Z_{(cm)}$. The figure shows corresponding orientations for the Y-gyro unit and the Z-gyro unit.

In order to utilize the output signals from an arrangement of gyro units like that of Fig. 6, servodrives must be mounted so that stabilizing torques may be applied between the successive gimbals of the system. In Fig. 6, one drive is shown between the middle gimbal and the inner gimbal; a second drive operates between the outer gimbal and the middle gimbal; and a third drive acts between the base and the outer gimbal.

7. RESOLUTION OF GYRO UNIT OUTPUT SIGNALS FOR OPERATION OF THE DRIVES IN A THREE-DEGREE-OF-FREEDOM GIMBAL STABILIZATION SYSTEM

Inspection of Fig. 6 immediately shows that, with the exception of the inner gimbal drive, the output signals from the gyro units cannot be directly used as inputs for the servodrives. The drive about the z-axis of the controlled member in combination with the Z-gyro unit represents a simple one-axis situation like that illustrated in Fig. 5. The output from this gyro unit can be directly connected to the Z-servodrive input. On the other hand, the X- and Y-gyro unit output signals must be properly divided between the middle gimbal drive and the outer gimbal drive. The nature of this division of signals and the equations that must be fulfilled by the operation of a signal distribution system are developed in Derivation Summary 1.

The first step of the development described in Derivation Summary 1 is to assume an angular velocity of the controlled member with respect to the inertial space reference

position for this member that would exist with the outputs from all three gyro units simultaneously at their null levels and then to resolve this angular velocity into components along the controlled member axes, $X_{(cm)}$, $Y_{(cm)}$, and $Z_{(cm)}$. For stabilization, the three servodrives would have to nullify these three angular velocity components on the basis of the three signals from the gyro units that would be proportional to the angular velocity component about the three controlled member axes. The procedure outlined in Derivation Summary 1 for finding the required servodrive angular velocity components is to project the controlled member components onto the gimbal drive axes. Trigonometric equations for the drive axes angular velocity components are given as Eqs. (1) through (4).

The angular deviation components of the controlled member with respect to the reference orientation that develop during any time interval can be found by integrations of the angular velocity components. When these integrations are started from an instant when the controlled member is at rest in the reference orientation and attention is restricted to the short time intervals required for any servodrives to overcome a deviation in one direction, all the angles involved remain small. Because of this fact, deviation angles may be transferred from controlled member axes to servodrive axes by projections similar to those used for angular velocity components.

The equations corresponding to these projections are given in Derivation Summary 1 for the correction angle components, which represent the angles through which the controlled member would have to be rotated in order to bring it into coincidence with its reference orientation as determined by null level signals from the gyro units. Fig. 6 includes representations of resolvers,

mounted so that they receive angular displacements between the gimbals on either side of the inner gimbal drive and the middle gimbal drive. These resolvers are used to distribute the X- and Y-gyro output signals to the middle and outer gimbal drives in accordance with the relationships developed in Derivation Summary 1.

8. FUNCTIONAL DIAGRAM FOR THREE-DEGREE-OF-FREEDOM GIMBAL STABILIZATION SYSTEM

Fig. 7 gives a functional diagram for the three-degree-of-freedom stabilization system represented in Fig. 6. The outer gimbal is mounted on the base with angular freedom about the $X_{(og)}$ -axis. The middle gimbal is similarly carried by the outer gimbal with angular freedom about the $Y_{(og)}$ -axis, and the inner gimbal, to which the controlled member is rigidly attached, is carried by the middle gimbal with angular freedom about the $Z_{(mg)}$ -axis.

The outer gimbal drive gear train and the outer gimbal drive motor have their housings mounted on the base and apply torque to the outer gimbal. The outer gimbal secant multiplier (Refs. 10 and 11) receives the angular position of the middle gimbal with respect to the outer gimbal and multiplies an electrical signal input by the secant of the relative gimbal angle. The middle gimbal drive motor and gear train combination applies torque to the middle gimbal, with the outer gimbal providing the necessary reaction. The inner gimbal drive motor and gear train combination performs a similar function for the inner gimbal with respect to the middle gimbal. The middle gimbal resolver is of the conventional synchro type and receives the angle of the inner gimbal with respect to the outer gimbal and multiplies each of two electrical input signals by the sine and cosine of the relative gimbal angle respectively.

The x-axis gyro unit, y-axis gyro unit, and z-axis gyro unit are mounted on the controlled member with the orientations shown in Fig. 6. The electrical output signal from the z-axis gyro unit is used as the feedback input for the z-axis drive power control system, which supplies voltage variations to the inner gimbal drive motor that determine the torque output of this motor. This drive power control system consists of five components connected in series. The preamplifier receives the z-axis gyro output signal voltage and changes it to a form suitable as the input for the demodulator. The output of the demodulator is a direct-current voltage input for the modifier, which provides filtering action and introduces the phase changes necessary to give dynamic stabilization for the inner gimbal drive servosystem loop.

The possible actions of electrical modifying circuits are discussed in many books dealing with the theory of electrical networks used in feedback systems (Ref. 12). The theory of servomechanisms, which is a special part of feedback theory, is treated in many books (Refs 13 to 18) that may be used as sources of information for designing the electrical components associated with gimbal drives. In Fig. 7, the modifier represents an arrangement of electrical components to introduce the dynamical effects necessary for a satisfactorily working servodrive system. The output from the modifier is the input for the remodulator, which, in turn, gives an output that acts to generate the voltage input for the drive amplifier. This amplifier raises the power level and provides an output current that causes the inner gimbal drive motor to apply stabilizing torque to the inner gimbal.

The feedback chains for the middle gimbal drive and outer gimbal drive follow the pattern of the closed chain for the inner gimbal drive except for the differences

required by the trigonometric transformations that are necessary to combine the x-axis and y-axis gyro output signals in accordance with the results of Derivation Summary 1. These signal combinations insure that the middle gimbal servodrive system and the outer gimbal servodrive system will operate with loop gains that are not affected by relative positions of the gimbals as long as the relative angle of the middle gimbal with respect to the outer gimbal remains somewhat less than 90 degrees.

9. STABILIZATION PERFORMANCE EQUATION FOR INNER GIMBAL DRIVE SYSTEM

The controlled member tends to maintain a fixed orientation with respect to inertial space because of its own inertia. It can also be demonstrated that no torques need be applied to the gimbals to maintain the controlled member in a fixed orientation in the presence of angular motion of the base, providing that the $Z_{(cm)}$ - and $Z_{(b)}$ - axes are essentially parallel and that $I_{(og)X} = I_{(og)Y} = 1/2 I_{(og)Z}$. (This implies that the outer gimbal is essentially a ring.) Under these conditions, the disturbing torques that must be canceled by the gimbal drive motor fall into the following classes:

- (1) Torque components due to non-viscous friction in the bearings.
- (2) Viscous friction torque components due to bearings and armature reaction torques from the motor.
- (3) Torque components due to linear accelerations acting on unbalanced gimbals.
- (4) Torque components required to accelerate the drive gear trains when the base is subjected to angular accelerations.

The performance equation for the stabilization of the controlled member by the inner gimbal drive system is given in Equation Summary 1. If, in addition to the previously stated assumptions that the $Z_{(cm)}$ - and $Z_{(og)}$ -axes are parallel and that $I_{(og)}X$, $I_{(og)}Y$, and $1/2 I_{(og)}Z$ are equal, it is assumed for convenience that the x-, y-, and z-axes of the various gimbals are close to being parallel, then the performance equations for the middle gimbal and outer gimbal drive systems may be obtained by cyclic permutation of the subscripts on the various symbols to identify the axes and the gimbal drives involved. The gyro unit output signals for the $X_{(cm)}$ - and $Y_{(cm)}$ -axes must be combined by the trigonometric relationships of Derivation Summary 1, and the gimbal moments of inertia must be introduced as required by different gimbal orientations.

Eq. (1) of Equation Summary 1 contains terms for the inertial effects the damping effects, and the electromagnetic effects that combine to produce the resultant torque which determines the rotational motion of the controlled member. Interference effects are taken into account by a separate term representing the resultant undesirable torque component from nonviscous bearing friction, sticking of mechanical parts, the action of linear acceleration and gravity on unbalanced masses in the controlled member and gimbals, and torques that may act on the system from any other sources.

Eq. (2) is the form taken by Eq. (1) under conditions in which rotational velocities and accelerations are so small that all the terms containing p as a factor are negligible in comparison to the angular displacement term on the left hand side of Eq. (1) and the interference torque term on the right hand side. In this situation, dynamic actions have no appreciable effect, and the frequency functions associated with the performance functions become equal to unity. This means

that the performance functions reduce to sensitivities. The resultant expression that appears as Eq. (2) shows that the ratio of the controlled member angle to the interference torque is equal to the reciprocal of the product of the sensitivities of the operating components of the system. This expression is, in effect, the elastic coefficient of the system in resisting torque applied in the direction of taking the controlled member away from its inertial space reference position.

Eq. (3) is the form taken by Eq. (1) when the forcing motion of the middle gimbal has a steady-state sinusoidal shape of such high frequency that the terms of Eq. (1) depending on inertial effects, i.e., the terms containing p^2 as a factor, are much larger than all the other terms. In this case, Eq. (1) reduces to a ratio of controlled member displacement to middle gimbal displacement that depends only on factors that contain moments of inertia. The numerator is the product of the gear train sensitivity for controlled member angle to rotor angle, a factor that contains this sensitivity minus unity, and the moment of inertia of the rotor. The denominator is the sum of the controlled member moment of inertia and the rotor inertia multiplied by the square of the gear train sensitivity. Eq. (3) means that at high frequencies the controlled member amplitude is always a constant fraction of the amplitude of the middle gimbal motion.

The responses of the controlled member of an illustrative gear-driven stabilization system to sinusoidal interfering torques and sinusoidal base displacements over a range of forcing frequencies are shown by the two nondimensional logarithmic scale curves of Fig. 8. The curve representing the response of the controlled member to interfering torques has an ordinate that is proportional to the constant low frequency stiffness of the servo up to the resonant frequency region of the system. In the region of the resonant

frequency, the low frequency stiffness due to the servodrive and the controlled member and gear train inertia interact to produce a peak in the curve. Above this frequency region, the curve drops off with a slope having a magnitude of two.

The curve representing the response of the controlled member to sinusoidal middle gimbal motion rises as the frequency increases up to the resonant frequency region. Above this region, the curve corresponds to a constant response amplitude ratio. The magnitude of the response amplitude is determined by the ratio of the effective gear train and rotor inertia to the sum of the controlled member inertia and the effective rotor and gear train inertia.

In practice, the overall response of the controlled member in a stabilization system is determined by the resultant of the effects represented by the two curves of Fig. 8. For a direct-drive servo, there is substantially no controlled member response due to rotation of the base; the only motion of the controlled member is that produced by the interfering torques. Examination of this curve clearly indicates that, with a direct-drive servo system, the servo need operate effectively only up to a frequency sufficiently high to insure adequate stabilization by the controlled member inertia. This reduces the servo problem from one of maintaining high stiffness over the entire frequency range to one of maintaining a high stiffness in only the low frequency range.

REFERENCES

1. Draper, C. S., McKay, Walter, and Lees, Sidney, "Instrument Engineering," McGraw-Hill Book Company, Inc., New York, Vol. I, 1952; Vol. II, 1953; Vol. III, Part 1, 1955; Vol. III, Part 2, being written.
2. Draper, C. S., and Woodbury, R. B., "Gyroscopic Apparatus," U. S. Patent 2,752,792, Application date March 22, 1951.
3. Draper, C. S., Hutzenlaub, J. F., and Woodbury, R. B., "Gyroscopic Apparatus," U. S. Patent 2,752,793, Application date March 22, 1951.
4. Draper, C. S., Wrigley, Walter, and Grohe, L. R., "The Floating Integrating Gyro and Its Application to Geometrical Stabilization Problems on Moving Bases," Sherman M. Fairchild Publication Fund Paper No. FF-13, Institute of the Aeronautical Sciences, New York, January 1955.
5. Draper, C. S., Wrigley, Walter, and Grohe, L. R., "The Floating Integrating Gyro and Its Application to Geometrical Stabilization Problems on Moving Bases," Aeronautical Engineering Review, Volume 15, No. 6, June 1956.
6. Draper, C. S., "Gyroscopic Apparatus," U. S. Patent 2,752,790, Application date August 2, 1951.
7. Jarosh, J. J., Haskell, C. A., and Dunnell, W. W., Jr., "Gyroscopic Apparatus," U. S. Patent 2,752,791, Application date February 9, 1951.

8. Ref. 1, Vol. III, Part 1, Figure 35-5.
9. Gardner, M. F., and Barnes, J. L., "Transients in Linear Systems," John Wiley & Sons, Inc., New York, 1942.
10. Greenwood, I. A., Jr., Holdam, J. V., Jr., and Macrae, Duncan, Jr., (editors), "Electronic Instruments," Radiation Laboratory Series, 21, McGraw-Hill Book Company, Inc. New York, 1948.
11. Ahrendt, W. R., "Servomechanism Practice," McGraw-Hill Book Company, Inc., New York, 1954.
12. Bode, H. W., "Network Analysis and Feedback Amplifier Design," D. Van Nostrand Company, Inc., New York, 1945.
13. Mac Coll, L. A., "Fundamental Theory of Servomechanisms," D. VanNostrand Company, Inc., New York, 1945.
14. Brown G.S., and Campbell, D.P., "Principles of Servomechanisms," John Wiley & Sons, Inc., New York, 1948.
15. James, H. M., Nichols, N. B., and Phillips, R. S., (editors), "Theory of Servomechanisms," Radiation Laboratory Series, 25, McGraw-Hill Book Company, Inc., New York, 1947.
16. Ahrendt, W. R., and Taplin, J. F., "Automatic Feedback Control," McGraw-Hill Book Company, Inc., New York, 1951.
17. Chestnut, Harold, and Mayer, R. W., "Servomechanisms and Regulating System Design," Vol. I-II, John Wiley & Sons, Inc., New York, 1951, 1955.
18. Truxal, J. G., "Automatic Feedback Control System Synthesis," McGraw-Hill Book Company, Inc., New York, 1955.

Based on the development given in Derivation Summaries of Reference⁴ with mechanical and electrical inaccuracy producing effects considered as negligible, the performance equation of the integrating gyro unit may be written in the form

$$e_{(gu)} = [PF]_{(gu)} | W; e | \left[W[I - (ca)](IA) - (SR)_{(gu)} | i; W | (ref) i_{(in)}(gu) - (SR)_{(gu)} | \ddot{A}; W | (ref) \ddot{A}[I - (ca)](OA) \right] \quad (1)$$

where

$e_{(gu)}$ = gyro unit output voltage

$[PF]_{(gu)} | W; e |$ = angular velocity input - voltage output performance function* of the gyro unit

$W[I - (ca)](IA)$ = angular velocity of gyro unit case about the input axis with respect to inertial space

$i_{(in)}(gu)$ = gyro unit input current (input current to the torque generator)

$\ddot{A}[I - (ca)](OA)$ = angular acceleration of gyro unit case about the output axis with respect to inertial space (Note: $\dot{} = d/dt$)

$$(SR)_{(gu)} | i; W | (ref) = \left[\frac{W[I - (ca)](IA)}{i_{(in)}(gu)} \right] (\dot{e} = 0) \quad \begin{aligned} &= \text{current - angular velocity reference sensitivity ratio magnitude of the gyro unit} \quad (2) \\ &= \text{ratio of angular velocity input magnitude to input current magnitude required to hold output voltage constant} \end{aligned}$$

$$(SR)_{(gu)} | \ddot{A}; W | = \left[\frac{W[I - (ca)](IA)}{\ddot{A}[I - (ca)](OA)} \right] (\dot{e} = 0) \quad \begin{aligned} &= \text{output axis angular acceleration - angular velocity reference sensitivity ratio of the gyro unit} \quad (3) \\ &= \text{ratio of angular velocity input magnitude to angular acceleration about output axis required to hold output voltage constant} \end{aligned}$$

$$[PF]_{(gu)} | W; e | = \frac{S_{(gu)} | W; \dot{e} | (ref)}{p[1 + p(CT)_{(gu)}]} \quad (4)$$

$$S_{(gu)} | W; \dot{e} | (ref) = \left[\frac{\dot{e}_{(gu)}}{W[I - (ca)](IA)} \right] \quad \begin{aligned} &\text{(under static conditions giving reference value of ratio)} \\ &= \text{angular velocity - voltage rate reference sensitivity** magnitude of the gyro unit} \end{aligned}$$

*A performance function may be associated with an ordinary differential equation form of performance equation so that it represents the effect of the differential equation in establishing relationships between the independent variable and the dependent variable. The performance function is particularly useful because it reduces differential equation manipulations to the processes of algebra. For example, a typical differential equation $\tau \dot{v} + v = \sigma u$ has a performance function $[PF]$, where $v = [PF]u$; see Draper, McKay and Lees, *Instrument Engineering* [24], Vol. II, Chapter 17.

**See footnote on next page.

Information Summary 1. Performance equation for a single-axis integrating gyro unit.
Illustrative summary of performance function definitions, conventions, and notation.
(Page 1 of 3.)

Note that

$$S_{(gu)}[W; \dot{e}]_{(ref)} = S_{(gu)}[A; e]_{(ref)} = \left[\frac{e_{gu}}{A[I - (ca)](IA)} \right] \begin{matrix} \text{(under static conditions} \\ \text{giving reference value} \\ \text{of ratio)} \end{matrix} = \begin{matrix} \text{angle - voltage refer-} \\ \text{ence sensitivity magni-} \\ \text{tude of the gyro unit} \end{matrix}$$

$$(CT)_{(gu)} = \frac{I_{(gim)}(OA)}{S_{(vs)}[\dot{A}; M]} \quad \text{characteristic time of gyro unit}$$

$$I_{(gim)}(OA) \quad \text{moment of inertia of the gimbal float about the output axis}$$

$$S_{(vd)}[\dot{A}; M] = \frac{M_{(vd)}}{\dot{A}[(ca) - (gim)](OA)} = \frac{\text{viscous damper torque acting on gimbal}}{\text{angular velocity of gimbal with respect to case about output axis}}$$

$$= \text{angular velocity input - torque output sensitivity of viscous damper}$$

When attention is restricted to steady-state sinusoidal changes,

$$e_{(gu)} = e_{(gu)a} e^{j\omega_f t}; \quad W[I - (ca)](IA) = W[I - (ca)](IA)a e^{j\omega_f t}; \dots$$

$$p = \frac{d}{dt} = j\omega_f = j2\pi n_f; \quad \omega_f = 2\pi n_f; \quad n_f = \text{forcing frequency}$$

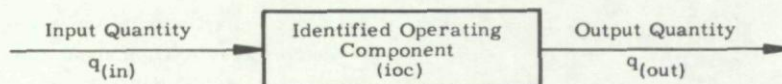
$$T_f = \frac{1}{n_f} = \text{forcing period}; \quad n_f = (FF)$$

When

$$W[I - (ca)](IA) \text{ is written as } p A[I - (ca)](IA) = j2\pi n_f A[I - (ca)](IA) \quad (5)$$

$$\ddot{A}[I - (ca)](OA) \text{ is written as } p^2 A[I - (ca)](OA) = - (2\pi n_f)^2 A[I - (ca)](OA) \quad (6)$$

**The generalized definition of the concept of sensitivity for an operating component is associated with the change of the output caused by a change in an input to an operating component.



$$S_{(ioc)}[q_{(in)}; q_{(out)}] = \frac{dq_{(out)}}{dq_{(in)}} = \text{input quantity - output quantity sensitivity of the identified operating component}$$

When the input quantity - output quantity relationship is linear over the operating range considered,

$$S_{(ioc)}[q_{(in)}; q_{(out)}] = \frac{q_{(out)}}{q_{(in)}}$$

The symbol for the corresponding sensitivity magnitude is

$$S_{(ioc)}[q_{(in)}; q_{(out)}] = \left| S_{(ioc)}[q_{(in)}; q_{(out)}] \right| = \left| \frac{q_{(out)}}{q_{(in)}} \right| = \text{input quantity - output quantity sensitivity magnitude of the identified operating component}$$

Refer to Volume I of *Instrument Engineering* by Draper, McKay and Lees [24] for the generalized conventions used in defining concepts and forming symbols.

Information Summary 1. Performance equation for a single-axis integrating gyro unit.
Illustrative summary of performance function definitions, conventions, and notation.
(Page 2 of 3.)

and for the purposes of stabilization performance

$$i_{(in)(gu)} = 0$$

the performance equation may be written in the form

$$e_{(gu)} = \frac{S_{(gu)}|A;e|(ref)}{[1 + p(CT)_{(gu)}]} \left[1 - p(SR)_{(gu)}|\ddot{A};W|(ref) \frac{A_{[I-(ca)](OA)}}{A_{[I-(ca)](IA)}} \right] A_{[I-(ca)](IA)} \quad (7)$$

Define

$$[(FF)(SR)P]_{(gu)}|\ddot{A};W|(ref) = n_f(SR)_{(gu)}|\ddot{A};W|(ref) = \text{forcing frequency - angular acceleration - angular velocity sensitivity ratio product}$$

Noting that

$$A_{[I-(ca)](IA)} = \text{angle of case with respect to inertial space reference about input axis}$$

$$A_{[I-(ca)](OA)} = \text{angle of case with respect to inertial space reference about output axis}$$

Let

$$p \rightarrow j 2\pi n_f = j 2\pi \frac{1}{T_f} \quad \text{and} \quad (CT)_{(gu)} n_f = [(CT)(FF)P] = \text{characteristic time forcing frequency product} \quad (8)$$

$$e_{(gu)} = \frac{S_{(gu)}|A;e|(ref)}{1 + j 2\pi [(CT)(FF)P]} \left[1 - j 2\pi [(FF)(SR)P]_{(gu)}|\ddot{A};W|(ref) \frac{A_{[I-(ca)](OA)}}{A_{[I-(ca)](IA)}} \right] A_{[I-(ca)](IA)} \quad (9)$$

By definition

$$\begin{aligned} [PF]_{(gu)}|A;e| &= S_{(gu)}|A;e|(ref) \frac{1}{1 + j 2\pi [(CT)(FF)P]} = S_{(gu)}|A;e|(ref) \frac{1}{\sqrt{1 + [2\pi((CT)(FF)P)]^2}} e^{j(DRA)_{(gu)}(A;e)} \\ &= S_{(gu)}|A;e| e^{j(DRA)_{(gu)}(A;e)} \end{aligned} \quad (10)$$

$$S_{(gu)}|A;e| = \text{angle - voltage sensitivity of gyro unit} = S_{(gu)}|A;e|(ref) [S(RS)R]_{(gu)}|A;e| \quad (11)$$

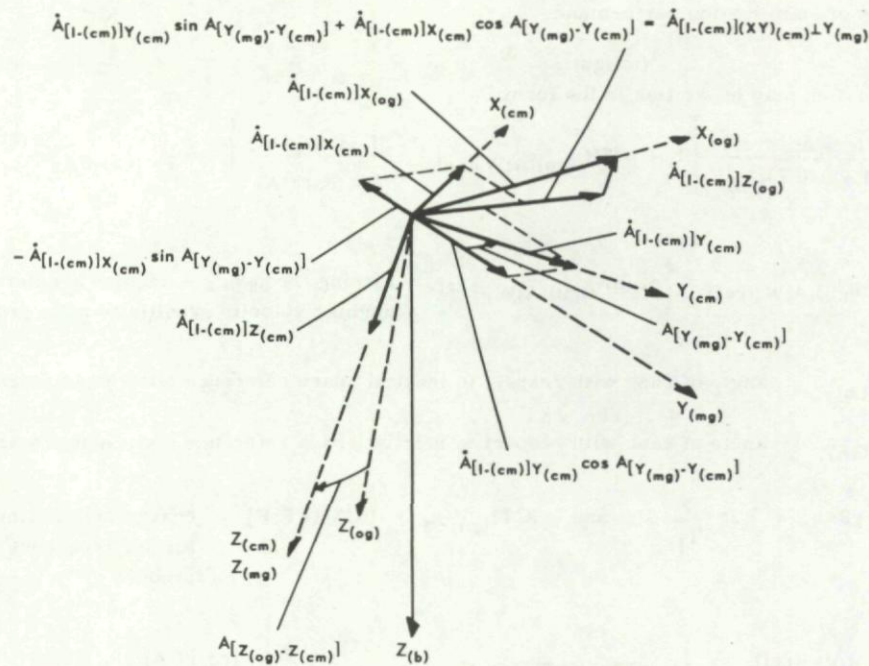
$$[S(RS)R]_{(gu)}|A;e| = \frac{S_{(gu)}|A;e|}{S_{(gu)}|A;e|(ref)} = \frac{1}{\sqrt{1 + [2\pi((CT)(FF)P)]^2}} = \text{sensitivity - reference sensitivity product of gyro unit} \quad (12)$$

Note

$$[S(RS)R]_{(gu)}|A;e| = (SR)_{(gu)}|A;e| = \text{angle - voltage dimensionless reference sensitivity of the gyro unit} \quad (13)$$

$$(DRA)_{(gu)}(A;e) = \text{angle - voltage dynamic response angle of gyro unit} \quad (14)$$

Information Summary 1. Performance equation for a single-axis integrating gyro unit.
Illustrative summary of performance function definitions, conventions, and notation.
(Page 3 of 3.)



$$\dot{A}_{[I-(cm)]Z(mg)} = \dot{A}_{[I-(cm)]Z(cm)} \quad (1)$$

$$\dot{A}_{[I-(cm)]Y(mg)} = \dot{A}_{[I-(cm)]Y(cm)} \cos A_{[Y(mg)-Y(cm)]} - \dot{A}_{[I-(cm)]X(cm)} \sin A_{[Y(mg)-Y(cm)]} \quad (2)$$

$$\begin{aligned} \dot{A}_{[I-(cm)]X(og)} = \sec A_{[Z(og)-Z(mg)]} & \left[\dot{A}_{[I-(cm)]Y(cm)} \sin A_{[Y(mg)-Y(cm)]} \right. \\ & \left. + \dot{A}_{[I-(cm)]X(cm)} \cos A_{[Y(mg)-Y(cm)]} \right] \quad (3) \end{aligned}$$

$$\begin{aligned} \dot{A}_{[I-(cm)]Z(og)} = \csc A_{[Z(og)-Z(mg)]} & \left[\dot{A}_{[I-(cm)]Y(cm)} \sin A_{[Y(mg)-Y(cm)]} \right. \\ & \left. + \dot{A}_{[I-(cm)]X(cm)} \cos A_{[Y(mg)-Y(cm)]} \right] \quad (4) \end{aligned}$$

a) Angular velocity components about the gimbal drive axes of the controlled member with respect to inertial space in terms of angular velocity components about the gyro unit input axes

Derivation Summary 1. Controlled member angular corrections about the gimbal drive axes in terms of correction components with respect to inertial space reference orientation about controlled member axis. (Page 1 of 4.)

Integrating Eqs. (1), (2), (3) and (4)

$$\int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Z_{(mg)}} dt = \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Z_{(cm)}} dt \quad (5)$$

$$\begin{aligned} \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Y_{(mg)}} dt &= \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Y_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} dt \\ &\quad - \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]X_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} dt \end{aligned} \quad (6)$$

$$\begin{aligned} \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]X_{(og)}} dt &= \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Y_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} \sec A_{[Z_{(og)} - Z_{(mg)}]} dt \\ &\quad + \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]X_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} \sec A_{[Z_{(og)} - Z_{(mg)}]} dt \end{aligned} \quad (7)$$

$$\begin{aligned} \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Z_{(og)}} dt &= \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]Y_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} \csc A_{[Z_{(og)} - Z_{(mg)}]} dt \\ &\quad + \int_{t=t_1}^{t=t_2} \dot{A}_{[I-(cm)]X_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} \csc A_{[Z_{(og)} - Z_{(mg)}]} dt \end{aligned} \quad (8)$$

For convenience the t_1 may be taken at an instant when the controlled member is coincident with its inertial space reference orientation and is not rotating with respect to this position. For the purposes of control system operation attention is directed toward the small angular deviations in controlled member orientation that occur in such short time intervals that the gimbal orientation angles $A_{[Y_{(mg)} - Y_{(cm)}]}$ and $A_{[Z_{(og)} - Z_{(mg)}]}$ may be treated as constants in carrying out the integrations of Eqs. (5), (6), (7) and (8). Under these conditions the integrations give

$$A_{[I-(cm)]Z_{(mg)}} = A_{[I-(cm)]Z_{(cm)}} \quad (9)$$

$$A_{[I-(cm)]Y_{(mg)}} = A_{[I-(cm)]Y_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} - A_{[I-(cm)]X_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} \quad (10)$$

Derivation Summary 1. Controlled member angular corrections about the gimbal drive axes in terms of correction components with respect to inertial space reference orientation about controlled member axis. (Page 2 of 4.)

$$\begin{aligned}
A_{[I-(cm)]X_{(og)}} &= A_{[I-(cm)]Y_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} \sec A_{[Z_{(og)} - Z_{(mg)}]} \\
&+ A_{[I-(cm)]X_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} \sec A_{[Z_{(og)} - Z_{(mg)}]}
\end{aligned}
\tag{11}$$

$$\begin{aligned}
A_{[I-(cm)]Z_{(og)}} &= A_{[I-(cm)]Y_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]} \csc A_{[Z_{(og)} - Z_{(mg)}]} \\
&+ A_{[I-(cm)]X_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} \csc A_{[Z_{(og)} - Z_{(mg)}]}
\end{aligned}
\tag{12}$$

b) Angular displacement components about gimbal drive axes of controlled member angles from the inertial space reference orientation

For the purposes of stabilization the controlled member displacement components with respect to the inertial reference orientation are deviation angles to be corrected by operation of the stabilization system.

In general, corrections are equal in magnitude and opposite in sign to the corresponding deviations. For example,

$$\begin{aligned}
(D)A_{[I-(cm)]Z_{(mg)}} &= \text{controlled member deviation about the middle gimbal Z-axis with respect to the inertial space reference orientation} \\
-(D)A_{[I-(cm)]Z_{(mg)}} &= (C)A_{[I-(cm)]Z_{(mg)}} = \text{controlled member correction angle about the middle gimbal Z-axis}
\end{aligned}
\tag{13}$$

where (D)() is the symbol for a deviation associated with the quantity represented by any symbol placed in the open parenthesis and

(C)() is the symbol for a correction associated with the quantity represented by any symbol placed in the open parenthesis.

From Eqs. (9) and (13)

$$(C)A_{[I-(cm)]Z_{(mg)}} = -(D)A_{[I-(cm)]Z_{(mg)}} = (C)A_{[I-(cm)]Z_{(cm)}} \tag{14}$$

Similarly, from Eqs. (10), (11), (12) and (13)

$$\begin{aligned}
(C)A_{[I-(cm)]Y_{(mg)}} &= (C)A_{[I-(cm)]Y_{(cm)}} \cos A_{[Y_{(mg)} - Y_{(cm)}]} \\
&- (C)A_{[I-(cm)]X_{(cm)}} \sin A_{[Y_{(mg)} - Y_{(cm)}]}
\end{aligned}
\tag{15}$$

Derivation Summary 1. Controlled member angular corrections about the gimbal drive axes in terms of correction components with respect to inertial space reference orientation about controlled member axis. (Page 3 of 4.)

$$\begin{aligned}
 (C)A[I - (cm)]X_{(og)} &= (C)A[I - (cm)]Y_{(cm)} \sin A[Y_{(mg)} - Y_{(cm)}] \sec A[Z_{(og)} - Z_{(mg)}] \\
 &+ (C)A[I - (cm)]X_{(cm)} \cos A[Y_{(mg)} - Y_{(cm)}] \sec A[Z_{(og)} - Z_{(mg)}]
 \end{aligned}
 \tag{16}$$

$$\begin{aligned}
 (C)A[I - (cm)]Z_{(og)} &= (C)A[I - (cm)]Y_{(cm)} \sin A[Y_{(mg)} - Y_{(cm)}] \csc A[Z_{(og)} - Z_{(mg)}] \\
 &+ (C)A[I - (cm)]X_{(cm)} \cos A[Y_{(mg)} - Y_{(cm)}] \csc A[Z_{(og)} - Z_{(mg)}]
 \end{aligned}
 \tag{17}$$

c) Corrections for controlled member deviations from the reference orientation

Derivation Summary 1. Controlled member angular corrections about the gimbal drive axes in terms of correction components with respect to inertial space reference orientation about controlled member axis. (Page 4 of 4.)

The performance equation of the inner gimbal stabilization drive may be written as

$$\begin{aligned}
 & \left\{ \left[I_{(cm)} Z_{(cm)} + S_{(dgt)}^2 [A_{(cm)}; A_{(rot)}]_{(ig)} I_{(rot)(dm)(ig)} \right] p^2 \right. \\
 & + \left[\left(S_{(dm)} \dot{A}; M \right)_{(ig)} + C_{(rot)(dm)(ig)} \right] S_{(dgt)}^2 [A_{(cm)}; A_{(rot)}]_{(ig)} + C_{(cm)} Z \Big] p \\
 & + S_{(dm)} \dot{i}; M \Big]_{(ig)} [PF]_{(dpcs)} [e; i]_{(ig)} [PF]_{(gu)} [A; e] Z S_{(dgt)} [A_{(cm)}; A_{(rot)}]_{(ig)} \Big\} A_{[I - (cm)] Z_{(cm)}} \\
 & = M_{(intfr)(cm)(ig)} + \left\{ p \left[\left(S_{(dm)} \dot{A}; M \right)_{(ig)} + C_{(rot)(dm)(ig)} \right] S_{(dgt)}^2 [A_{(cm)}; A_{(rot)}]_{(ig)} + C_{(cm)} Z \right] \\
 & + p^2 I_{(rot)(dm)(ig)} S_{(dgt)} [A_{(cm)}; A_{(rot)}]_{(ig)} \left[S_{(dgt)} [A_{(cm)}; A_{(rot)}]_{(ig)} - 1 \right] \Big\} A_{[I - (mg)] Z_{(cm)}}
 \end{aligned} \tag{1}$$

where

$A_{[I - (cm)] Z_{(cm)}}$ = angular displacement of the controlled member about the $Z_{(cm)}$ -axis with respect to the inertial space orientation for which the Z -axis gyro output signal has its null level

$I_{(cm)} Z$ = moment of inertia of controlled member about the $Z_{(cm)}$ -axis

$S_{(dgt)} [A_{(cm)}; A_{(rot)}]_{(ig)}$ = controlled member angle input - rotor angle output sensitivity of inner gimbal drive gear train (identical with the low speed to high speed gear ratio of the gear train)

$I_{(rot)(dm)(ig)}$ = moment of inertia of the inner gimbal drive motor about its axis of rotation. This includes a component that depends on the effective inertia of the gear train reduced to rotor speed

$S_{(dm)} \dot{A}; M \Big]_{(ig)}$ = angular velocity input - torque output sensitivity magnitude of drive motor. This sensitivity is due to reduction in motor torque by internal electrical effects as speed increases

$C_{(rot)(dm)(ig)}$ = coefficient of viscous damping acting on inner gimbal drive motor rotor - referred to rotor speed

$C_{(cm)} Z$ = coefficient of viscous damping acting on controlled member about $Z_{(cm)}$ -axis - referred to controlled member speed

$S_{(dm)} \dot{i}; M \Big]_{(ig)}$ = voltage input - torque output sensitivity magnitude of inner gimbal drive motor under zero speed condition

Equation Summary 1. Performance equation for stabilization of the controlled member by the inner gimbal drive system. (Page 1 of 2.)

$[PF]_{(dpcs)[e;i](ig)} = S_{(dpcs)[e;i](ref)(ig)} (FF)_{(dpcs)[e;i](ig)}$ = voltage input - current output performance function of the inner gimbal drive power control system

$S_{(dpcs)[e;i](ig)}$ = voltage input - current output reference sensitivity of the inner gimbal drive power control system

$(FF)_{(dpcs)[e;i](ig)}$ = voltage input - current output frequency function of the inner gimbal drive power control system. When frequency $\rightarrow 0$, $(FF)_{(dpcs)[e;i](ig)} \rightarrow 1$

$[PF]_{(gu)[A;e]Z} = S_{(gu)[A;e](ref)Z} (FF)_{(gu)[A;e]Z}$ = angle input - voltage output performance function of the Z-axis gyro unit

$S_{(gu)[A;e](ref)Z}$ = angle input - voltage output reference sensitivity of the Z-axis gyro unit

$(FF)_{(gu)[A;e]Z}$ = angle input - voltage output frequency function of the Z-axis gyro unit. When frequency $\rightarrow 0$, $(FF)_{(gu)[A;e]Z} \rightarrow 1$

$M_{(intfr)(cm)Z_{(cm)}}$ = interference torque acting on the controlled member about the $Z_{(cm)}$ -axis

$A_{[I - (mg)]Z_{(cm)}}$ = angular displacement of the middle gimbal about the $Z_{(cm)}$ -axis with respect to a reference orientation in inertial space in which the Y-axis of the gimbal is aligned with the Y-axis of the controlled member when the controlled member is in the reference orientation determined by the null level signal of the Z-axis gyro unit

- a) Performance equation for the inner gimbal stabilization loop when the only motions with respect to inertial space of the middle gimbal and the controlled member are about the $Z_{(cm)}$ -axis

When the system is stationary, i.e., at zero frequency and with $A_{[I - (mg)]Z_{(cm)}} = 0$

$$\frac{A_{[I - (cm)]Z_{(cm)}}}{M_{(intfr)(cm)(ig)}} = \frac{1}{S_{(dm)[i;M](ref)(ig)} S_{(dpcs)[e;i](ref)(ig)} S_{(gu)[A;e](ref)Z} S_{(dgt)[A_{(cm)};A_{(rot)}](ig)}} \quad (2)$$

= angular stiffness of stabilization control in radians per unit torque

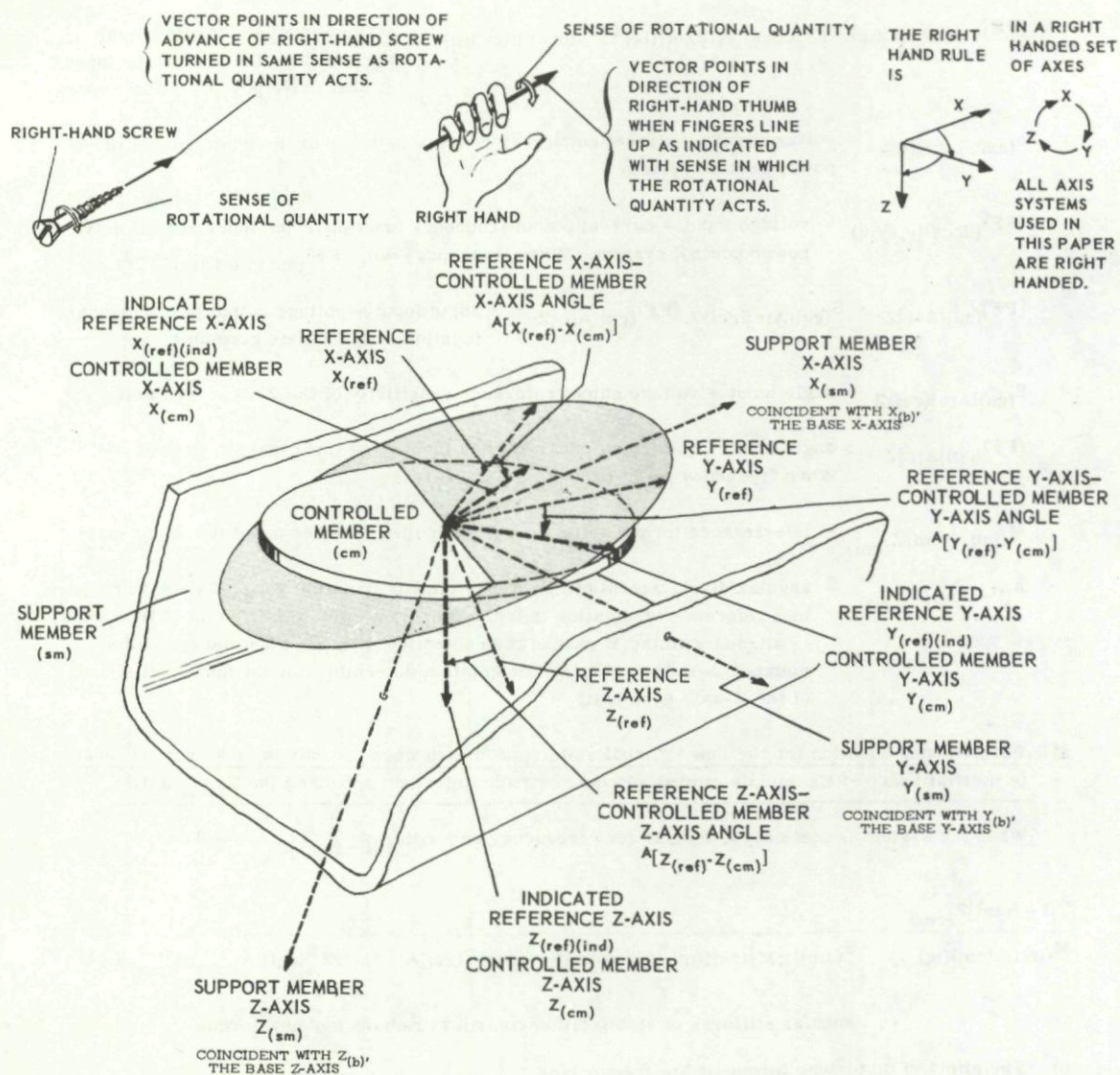
- b) The effect of disturbing torque at low frequencies

At very high frequencies - where p^2 terms are so great that all others are negligible

$$\frac{A_{[I - (cm)]Z_{(cm)}}}{A_{[I - (mg)]Z_{(cm)}}} = \frac{S_{(dgt)[A_{(cm)};A_{(rot)}](ig)} \left[S_{(dgt)[A_{(cm)};A_{(rot)}](ig)} - 1 \right] I_{(rot)(im)(ig)}}{I_{(cm)Z_{(cm)}} + S_{(dgt)[A_{(cm)};A_{(rot)}](ig)}^2 I_{(rot)(im)(ig)}} \quad (3)$$

- c) The effect of base motion interference at high frequencies

Equation Summary 1. Performance equation for stabilization of the controlled member by the inner gimbal drive system. (Page 2 of 2.)



NOTES: THE CONTROLLED MEMBER IS CARRIED BY A SYSTEM MOUNTED ON A BASE NOT SHOWN IN THIS DIAGRAM. THE BASE IS CARRIED BY THE SUPPORT MEMBER.

SUBSCRIPT LETTER ON ANY MAIN SYMBOL IDENTIFY THE ENTITY WITH WHICH THE QUANTITY REPRESENTED BY THE MAIN SYMBOL IS ASSOCIATED - THUS $X_{(cm)}$ IS THE X-AXIS ASSOCIATED WITH THE CONTROLLED MEMBER (SYMBOL (cm)).

THE SYMBOL $A_{[(ref)-(cmprd)]}$ REPRESENTS THE ANGLE A , MEASURED FROM THE REFERENCE DIRECTION ((ref) IN THE SUBSCRIPT) TO THE COMPARED DIRECTION ((cmprd) IN THE SUBSCRIPT)

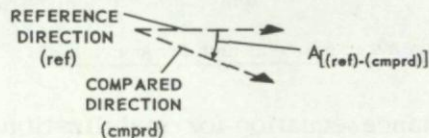
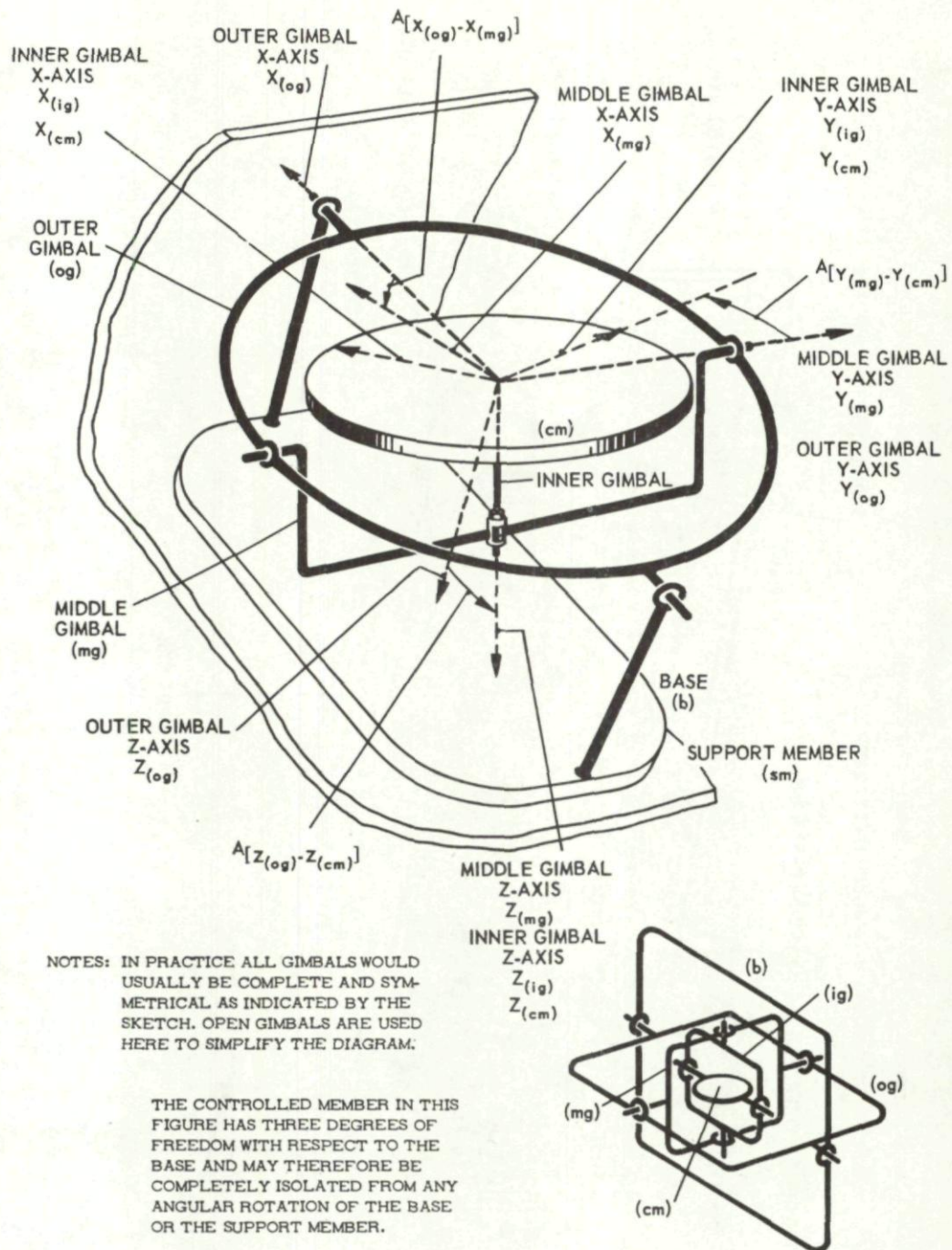
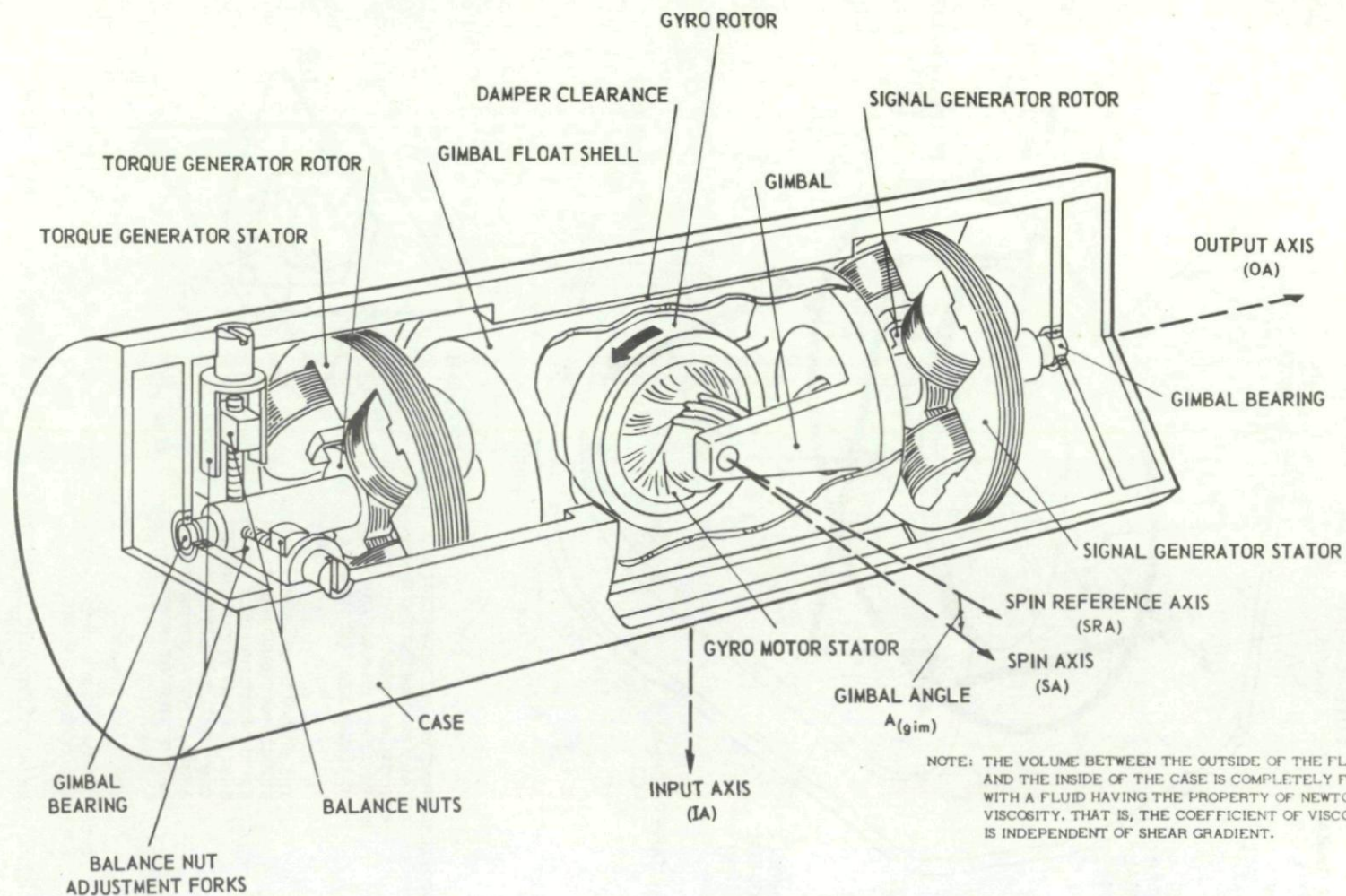


Fig. 1. Geometrical quantities associated with the basic problem of geometrical stabilization.



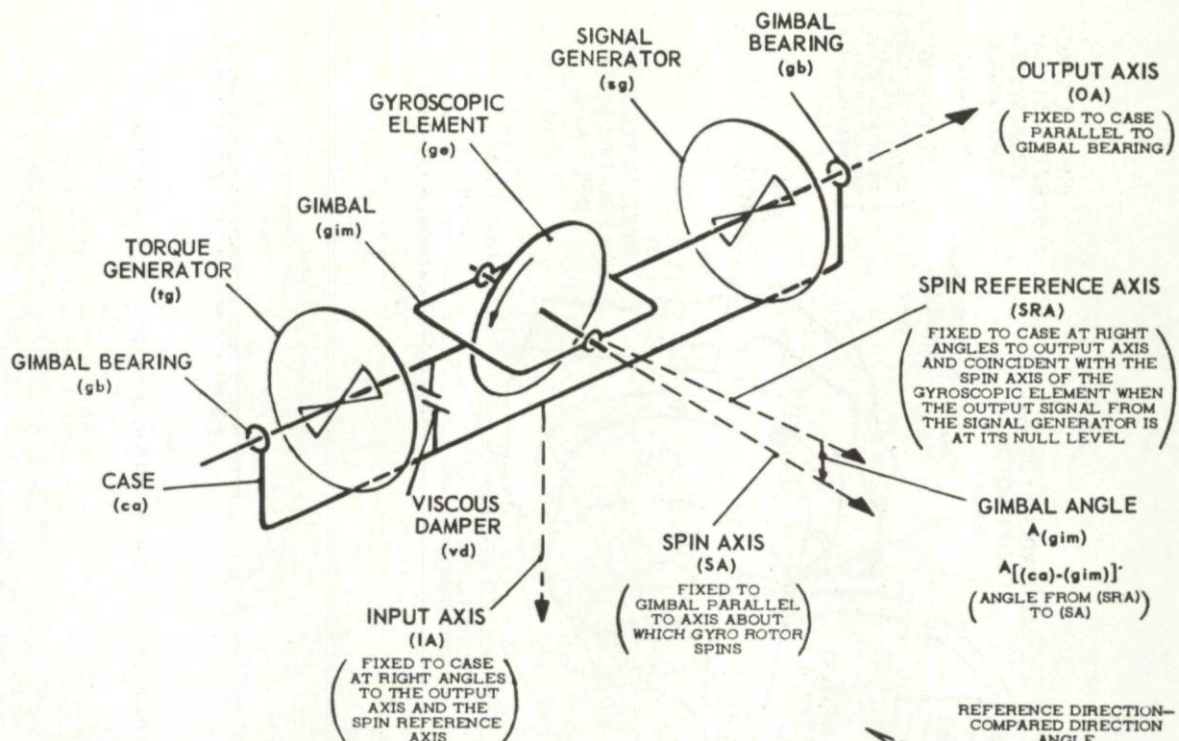
THIS DIAGRAM IS BASED ON FIG. 4 OF U. S. PATENT 2,752,792 AND FIG. 5 OF U. S. PATENT 2,752,793, DATED JULY 3, 1956.

Fig. 2. Line schematic diagram of gimbal system for stabilization.



THIS DIAGRAM IS BASED ON U. S. PATENT 2,752,791, FIG. 1 OF U. S. PATENT 2,752,790 AND FIG. 12 OF THE SHERMAN M. FAIRCHILD PUBLICATION FUND PAPER NO. FF-13, INSTITUTE OF THE AERONAUTICAL SCIENCES, NEW YORK, JANUARY 1955. (USED WITH PERMISSION)

Fig. 3. Pictorial diagram for the single-axis integrating gyro unit.



NOTES:

1. POSITIVE SENSES SHOWN BY THE ARROWS ARE CHOSEN SO THAT (IA), (SRA), AND (OA) FORM A RIGHT-HANDED SYSTEM.
2. THE GYRO UNIT TEMPERATURE CONTROL POWER IS SUPPLIED TO A MOUNTING BLOCK ADAPTED TO RECEIVE THE GYRO UNIT CASE. THE FLOW OF POWER IS CONTROLLED BY THE DAMPER TEMPERATURE SETTING.

3. THE SYMBOL $A[(ref)-(cmpd)]$ REPRESENTS THE ANGLE A MEASURED FROM THE REFERENCE DIRECTION ((ref) IN THE SUBSCRIPT) TO THE COMPARED DIRECTION ((cmpd) IN THE SUBSCRIPT).*

CASE - (ca) - THE STRUCTURE THAT GIVES SUPPORT FOR THE INTERNAL WORKING PARTS OF THE GYRO UNIT, ENCLOSES THE PARTS, AND CARRIES PROVISIONS FOR EXTERNAL CONNECTIONS OF ALL KINDS.

TORQUE GENERATOR - (tg) - COMPONENT FOR RECEIVING INPUT SIGNALS AND PRODUCING CORRESPONDING OUTPUT TORQUE APPLIED TO THE GIMBAL ABOUT THE OUTPUT AXIS.

DAMPER - (dmp) - SUBSYSTEM RECEIVING ANGULAR VELOCITY OF THE GIMBAL WITH RESPECT TO THE CASE AS ITS INPUT AND PRODUCING AS OUTPUT A RETARDING TORQUE ACTING ON THE GIMBAL ABOUT THE OUTPUT AXIS WITH A MAGNITUDE PROPORTIONAL TO THE MAGNITUDE OF THE ANGULAR VELOCITY OF THE GIMBAL WITH RESPECT TO THE CASE.

GYRO UNIT - (gu) - THIS ENTITY MADE UP OF THE COMPONENTS REPRESENTED IN THIS DIAGRAM AND ALL THE ADDITIONAL PARTS NECESSARY FOR A SINGLE PACKAGE TO CARRY OUT THE FUNCTIONS OF A GYRO UNIT.

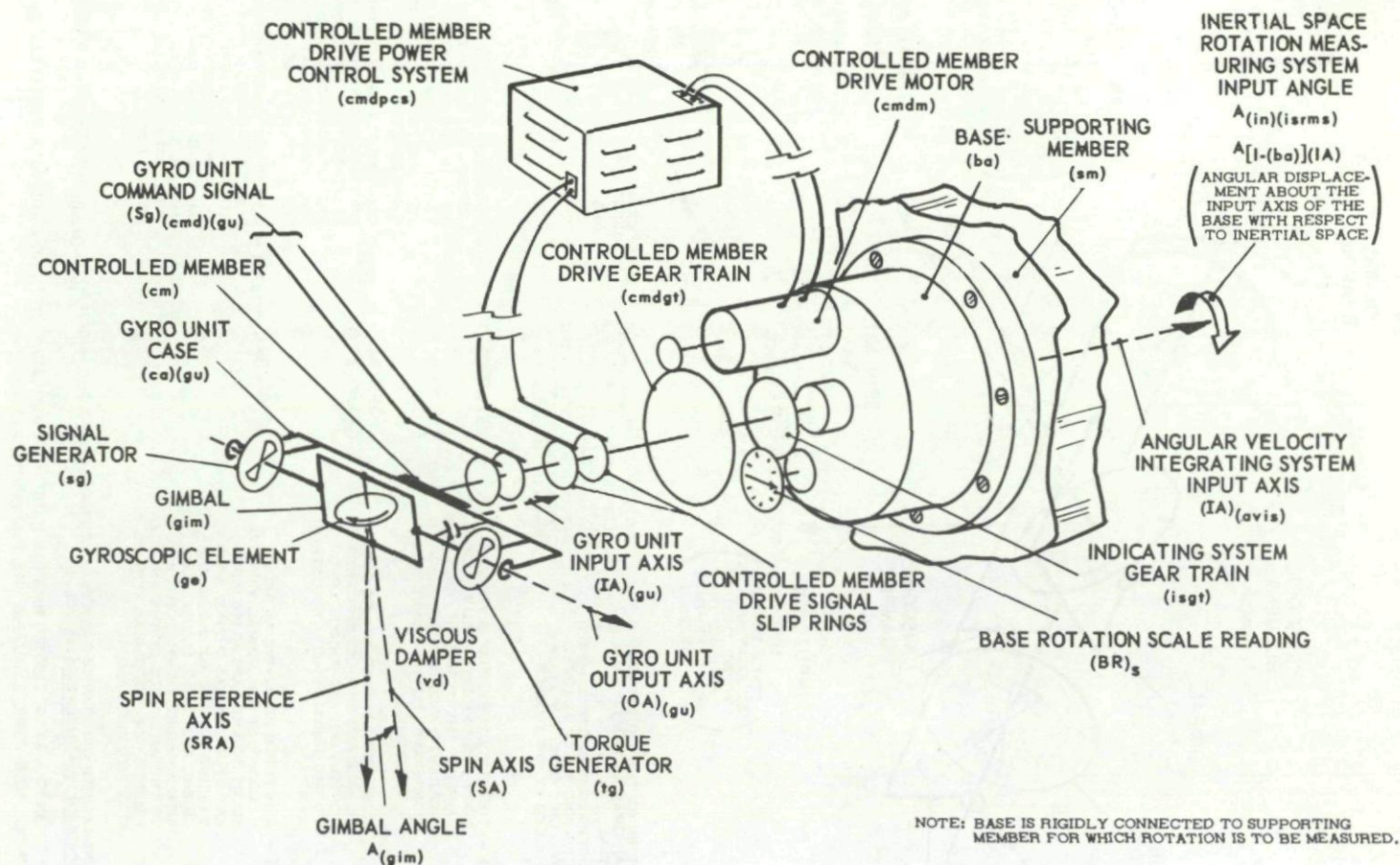
SIGNAL GENERATOR - (sg) - COMPONENT FOR RECEIVING THE ANGLE OF THE SPIN AXIS WITH RESPECT TO THE CASE AS INPUT AND PRODUCING A CORRESPONDING SIGNAL THAT SERVES AS THE OUTPUT SIGNAL FROM THE GYRO UNIT.

GIMBAL - (gim) - STRUCTURE CARRYING THE BEARINGS FOR THE SPINNING ROTOR OF THE GYROSCOPIC ELEMENT, ROTORS FOR THE TORQUE GENERATOR AND SIGNAL GENERATOR, PART OF THE DAMPER, FLOAT SEALS AND STRUCTURE, BALANCE ADJUSTMENTS, STOPS, PIVOTS, ETC.

* A DISCUSSION OF GENERALIZED CONVENTIONS FOR SELF-DEFINING SYMBOLS OF WHICH $A[(ref)-(cmpd)]$ IS AN EXAMPLE IS GIVEN BY DRAPER, MCKAY AND LEES IN INSTRUMENT ENGINEERING [24], VOL. I.

THIS DIAGRAM IS BASED ON FIG. 13 OF THE SHERMAN M. FAIRCHILD PUBLICATION FUND PAPER NO. FF-13, INSTITUTE OF THE AERONAUTICAL SCIENCES, NEW YORK, JANUARY 1955. (USED WITH PERMISSION)

Fig. 4. Line schematic for the single-axis integrating gyro unit.



THIS DIAGRAM IS BASED ON FIG. 14 OF THE SHERMAN M. FAIRCHILD PUBLICATION FUND PAPER NO. FF-13, INSTITUTE OF THE AERONAUTICAL SCIENCES, NEW YORK, JANUARY 1955. (USED WITH PERMISSION)

Fig. 5. Line schematic diagram of single-axis servo-driven controlled member with integrating gyro.

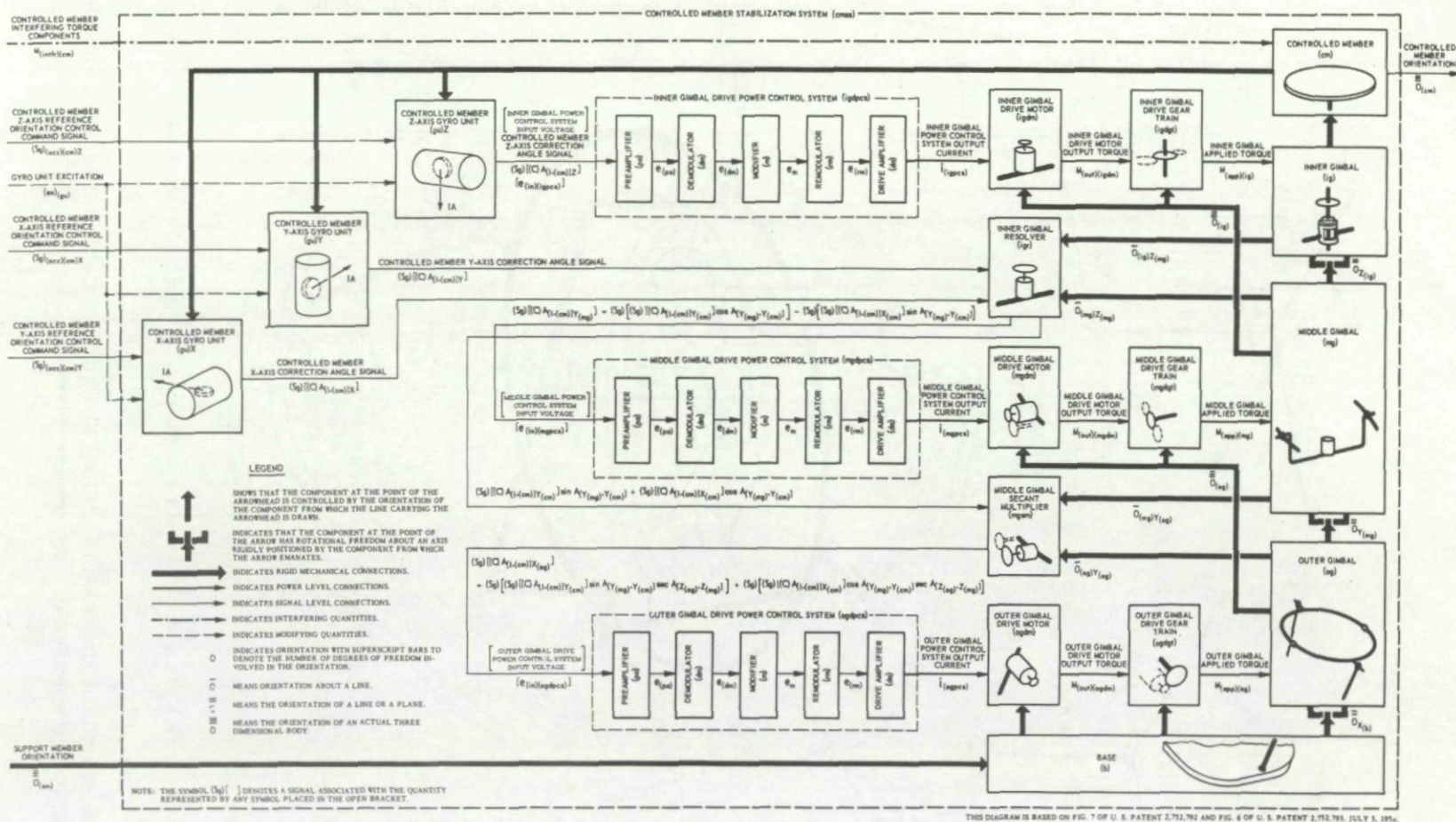
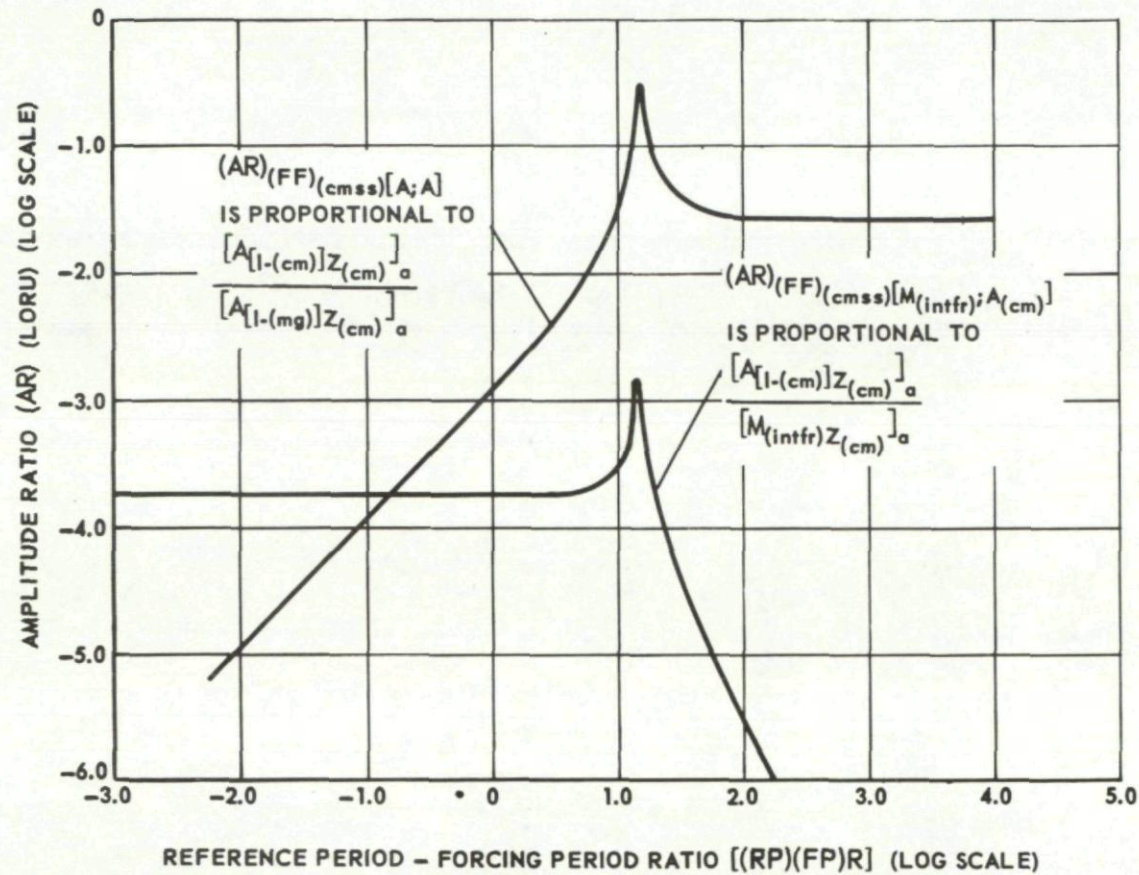


Fig. 7. Functional diagram for three-gimbal geometrical stabilization system based on servo drives actuated by three single-degree-of-freedom integrating gyro units.

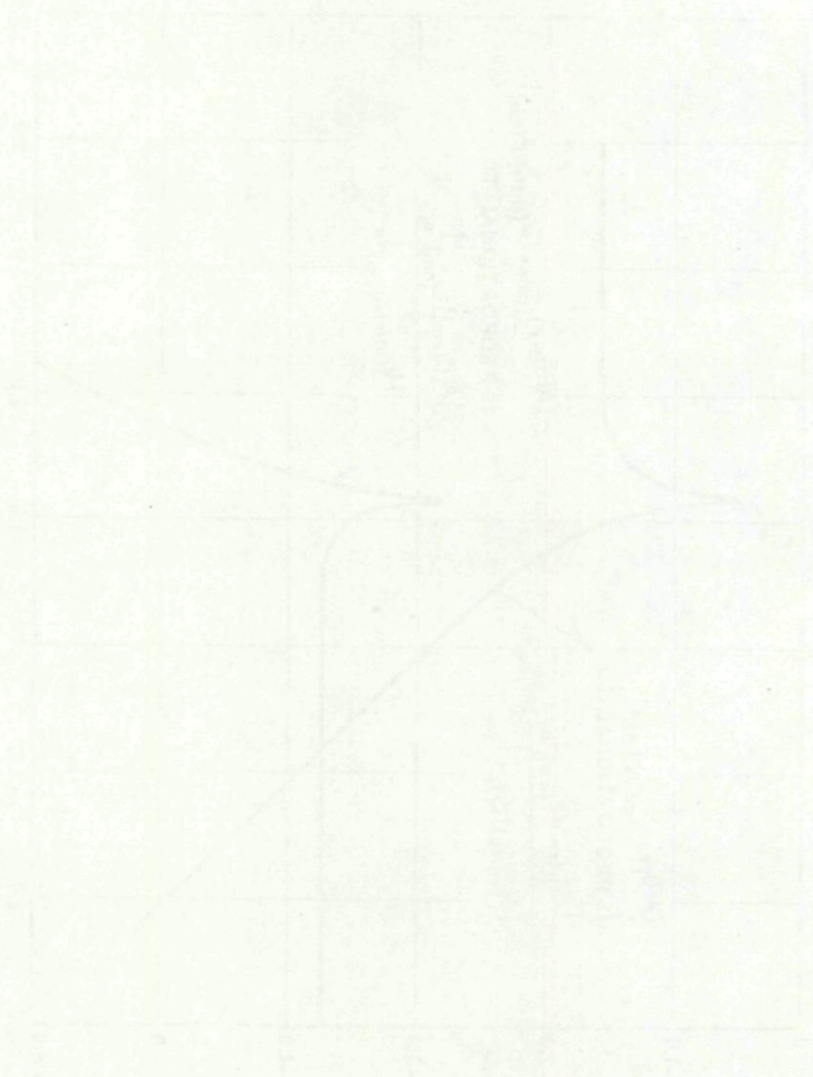
Fig. 8. Amplitude ratio variations of stabilization system for controlled member output produced by oscillatory interference torque and by oscillatory motion of middle gimbal.



$[A_{l-(cm)}Z_{(cm)}]_a$ = AMPLITUDE OF CONTROLLED MEMBER OSCILLATION ABOUT $Z_{(cm)}$ AXIS

$[A_{l-(mg)}Z_{(cm)}]_a$ = AMPLITUDE OF MIDDLE GIMBAL OSCILLATION ABOUT $Z_{(cm)}$ AXIS

$[M_{(intfr)}Z_{(cm)}]_a$ = AMPLITUDE OF INTERFERENCE TORQUE APPLIED TO CONTROLLED MEMBER ABOUT $Z_{(cm)}$ AXIS



THEORY OF THE CURVE

SAMPLED-DATA SYSTEMS

John R. Ragazzini*

SUMMARY

The basic theory, design, and application of sampled-data guidance and control systems are described. This description includes the mathematical characterization of the sampling process, use of the z -transform, stability of sampled systems, pulsed transfer functions, and a discussion of the advantages and disadvantages of this type of control.

SOMMAIRE

La théorie de base, l'étude et l'application des systèmes de contrôle et de gouverne par informations pulsées sont décrites. Cette description comprend l'étude: des caractéristiques mathématiques du procédé de pulsation, de l'utilisation de la transformation " z ", de la stabilité des systèmes pulsés, des fonctions de transfert des systèmes pulsés et finalement une discussion sur les avantages et les désavantages de ce type de contrôle.

1. INTRODUCTION

Sampled-data systems play an important role in guidance and control systems for missiles, processes, and aircraft. The characterizing feature of such systems is the fact that the signal data appear at one or more points of the system as a pulse sequence or as a sequence of numbers. Any intermittent element such as a time-shared data link, a digital computer, a scanning radar, or other data gathering transducer will cause the control system incorporating such an element to be a sampled-data system.

Sometimes sampled-data systems exist because one of the elements is unavoidably sampled. For instance, a scanning radar used for control will yield a fix on the target or controlled aircraft once each scanning period. Between such samples, data must be interpolated by some means or other. On the other hand, some systems are deliberately made sampled when they incorporate a digital computer as a controller. The computer can

accept numbers and compute steering commands periodically but not continuously. The increased sophistication of control which can result by use of a digital computer which may be required for data reduction in any case makes the design of an otherwise continuous system as a sampled-data system entirely desirable.

From the viewpoint of analysis and design, sampled-data systems present unique problems. Without going into detail at this point, it is evident that the sampling process reduces the information content of the function being sampled. For instance, if a scanning radar is used to determine the track of an airborne vehicle, the position is determined once each time the radar beam sweeps past the target. Between scans, no positive information is obtained so that if the target undertakes a turn or maneuver, this is not detected until the next and subsequent scans. Had the target been continuously irradiated by means of a tracking radar, the initiation of a maneuver would be immediately detected.

*Columbia University, New York, New York.

Another factor of importance to application of sampled systems to feedback control is the effect the process of sampling has on the stability of the system. Low-pass control systems tend to be unstable due to time lags in the loop. The introduction of sampling switches means that changes in the signal are not recognized until one or more sampling intervals later. Thus, continuous systems which are stable with a comfortable margin can become unstable if a sampling operation is introduced in some part of the loop. Theory which describes the performance of sampled-data systems both from the viewpoint of stability and response has been developed in various countries as will be seen in the list of references. Subsequent sections will deal with the status of theory and applications to the analysis and synthesis of sampled-data feedback control systems.

2. ELEMENTS OF A SAMPLED-DATA SYSTEM

While there is no unique form of sampled-data system, all such systems contain common elements arranged in various configurations. These elements can be categorized as samplers, data holds, digital controllers, and continuous linear (or non-linear) plants. These elements will be described individually and their characteristics described both qualitatively and mathematically. To indicate where each of these elements may be found, a typical configuration is shown in Fig. 1 for illustrative purposes. As will be shown later, there are many other possible configurations.

The first element which will be considered is the sampler which is shown here as a mechanical switch only to symbolize the operation. Referring to Fig. 2, the continuous function which is being sampled is $e_1(t)$. This function is sampled for a very short time τ and the output of the switch is a

sequence of short pulses whose area is proportional to the value of the time function at integral multiples of the sampling period T . The pulse sequence which results is identified by the function $e^*(t)$. The resultant pulse sequence may be obtained mathematically by multiplying the continuous function $e(t)$ by a carrier signal consisting of a periodic sequence of pulses referred to here as $p(t)$. Thus,

$$e^*(t) = e(t) p(t). \quad (1)$$

Since $p(t)$ is a periodic sequence, it may be expressed as a Fourier series so that

$$e^*(t) = e(t) \sum_{k=-\infty}^{+\infty} C_k e^{j \frac{2\pi k}{T} t} \quad (2)$$

where C_k 's represent the Fourier coefficients of the various terms. It is noted here that if the sampling function consists of very short pulses whose duration time τ is negligible, the various Fourier coefficients C_k tend to be equal. This condition is recognized as the "impulse sampling" approximation used by most investigators to characterize the sampling operation. It may be pointed out here that this assumption is ideal but also that it fits the practical situation very well and is very commonly used in analytical treatments of sampled-data systems.

Transform methods play an important part in the analysis of linear systems so that it is profitable to examine the Fourier and Laplace transforms of sampled time functions. The Fourier transform of the sampled function given in Eq. (2) is

$$F^*(j\omega) = \sum_{k=-\infty}^{+\infty} C_k F(j\omega + j \frac{2\pi k}{T}) \quad (3)$$

where $F(j\omega)$ is the Fourier transform of the continuous time function before sampling and $F^*(j\omega)$ is the Fourier transform of the sampled function and k has only integral values. This result is useful for illustrating the band limiting effects of the sampling operation. The result is shown graphically in Fig. 3 where the spectrum of the sampled function $e^*(t)$ is plotted.

It is seen that the spectrum of the original signal prior to sampling is repeated periodically after sampling. If the pulse which samples the signal is very short, the magnitudes of the repeated spectra are almost the same magnitude. Thus, the sampled signal has a Fourier spectrum $F^*(\omega)$ which contains components with frequencies not contained in the original signal. In order to extract the original signal from the pulse train resulting from the sampling operation, a low-pass filter must be employed. If the spectra are finite, that is, if they terminate completely at some finite maximum frequency F , it follows that a perfect filter can conceivably separate and recover the original signal provided the sampling frequency is twice the maximum frequency F contained in the signal spectrum. This is the well-known Shannon sampling theorem and has significance in this application that if the sampling rate is too low relative to the frequencies contained in the signal being sampled, loss in information results due to the impossibility of completely separating out the signal spectrum by filtration. Thus, a good rule is that in practical situations, it should not be expected to design control systems having an effective frequency response much greater than one-quarter to one-tenth the frequency of the sampling operation.

It is common in sampled-data systems to utilize a sampling function which is very short so that it may be expressed by a sequence of delta functions whose areas are

equal to unity. This mathematical approximation is extremely useful because it is mathematically convenient and expedient. Expressed mathematically, $p(t)$ becomes

$$p(t) = \sum_{n=-\infty}^{+\infty} \delta(t - nT). \quad (4)$$

Thus, a signal $e(t)$ which is being sampled by a sampling function of this type is given by

$$e^*(t) = \sum_{n=0}^{\infty} e(nT) \delta(t - nT). \quad (5)$$

If the original signal function $e(t)$ has a Laplace transform $E(s)$, then the sampled function $e^*(t)$ has a Laplace transform $E^*(s)$ which is given by

$$E^*(s) = \sum_{n=0}^{+\infty} e(nT) e^{-nTs}. \quad (6)$$

It is fortunate that for most systems the infinite series can be expressed in closed form thus making mathematical operations relatively straightforward. For instance, if the function $e(t)$ is a step function, then $e(nT)$ is a sequence of unit values for all positive values of n . $E^*(s)$ becomes for this case

$$E^*(s) = \sum_{n=0}^{\infty} e^{-nTs} \quad (7)$$

which is equal to

$$E^*(s) = \frac{1}{1 - e^{-Ts}}. \quad (8)$$

Similarly, it is possible to express all sequences of this type for signals whose Laplace transforms are ratios of rational polynomials in s in closed form as a ratio of rational polynomials in e^{Ts} . For this reason, tables of transforms of impulse sampled transfer functions can be drafted. Because the transforms of the sampled signals always contain the complex frequency s in the form e^{Ts} , the latter is replaced by an auxiliary variable "z" which has been used since the early work in the field by Hurwicz (Ref. 1). Thus Eq. (8) can be rewritten as

$$E(z) = \frac{1}{1 - z^{-1}} \quad (9)$$

where z has been replaced e^{Ts} .

Transforms of sampled time functions are thus referred to as "z-transforms" and are written in terms of z rather than e^{Ts} . Table 1 contains a list of commonly used z-transforms. Refs. 2 and 11 give more complete tables. The use of closed transforms to describe the sampled functions at the output of the sampler is a powerful tool used to analyze and synthesize sampled-data systems.

Referring once again to the system shown in Fig. 1, it is seen that the sampled error function is applied to a digital controller which processes the pulse sequence and delivers a modified pulse sequence to the data hold. Much more will be said about the digital controller later and at this point it will suffice to state that a digital program is implemented by this controller such that the output sequence causes the system to perform in some prescribed manner overall.

The output of the numerical operations carried out by the digital controller is a sequence of short pulses, or stated differently, a sequence of numbers. Were these

short pulses to be applied directly to a plant which must be controlled, it is evident that the plant would be subjected to severe shocks. The "duty cycle" of the command signal would be too low for practical purposes. For instance, if the controlled plant were an electric motor, the armature would be subjected to a sequence of short pulses which would produce only a small average power or torque. In order to operate the plant more efficiently, it is necessary to smooth out the pulse sequence, or fill in the voids between samples with an extrapolated time function that approximates a continuous function prior to sampling.

The system that accomplishes this data reconstruction or extrapolation is known as a "data hold," "data extrapolator," or "desampling filter." The function of this device may best be described by stating that it attempts to reconstruct the original continuous time function from which the pulse sequence was derived. In general, it cannot be expected that the function be accurately reconstructed so that only an approximation will result.

The simplest type of data hold system is the "clamp" circuit which assumes that the value of the function during the sampling interval is equal to the value at the sampling instant at the start of the interval. Thus, the reconstructed function will appear as a "staircase" function as shown in Fig. 4. This is the type of reconstruction obtained by storing the latest pulse in a digital register until the onset of a new sample which refreshes the number held by the register to the new value. The reconstructed function contains spurious signal components which can be recognized as periodic with a frequency of the sampler and its harmonics. These spurious frequencies are collectively referred to as "ripple." One of the important problems in sampled-data control systems is to maintain the ripple within specifications.

A mathematical representation of the data hold is needed to understand its effect in control systems. As an illustration, the transfer function of the simple clamp circuit will be obtained. As stated previously, the clamp circuit retains the value of the function as that of the previous sample. It is assumed here that the impulse modulation representation of the sampling process is sufficiently accurate for analysis. In this case, the clamp circuit has an impulsive response which is as shown in Fig. 5. An impulse of unit area is applied to the clamp circuit whose output rises to a unit value and then drops off to zero at the end of the sampling interval to be reset to its new value upon application of the next sample.

It is seen that the broad pulse resulting from the application of an impulse to the clamp circuit can be composed by adding two unit step functions, a positive one initiated at zero time and the other negative step applied at a delay equal to the sampling interval. Thus the impulsive response of the clamp is given by

$$g_h(t) = u(t) - u(t - T) \quad (10)$$

where $u(t)$ represents the unit step function.

Obtaining the Laplace transform of the impulsive response yields the transfer function of the clamp circuit. This is

$$G_h(s) = \frac{1 - e^{-Ts}}{s} \quad (11)$$

Eq. 11 is useful in determining the effect of the clamp circuit in a complete control system. It is interesting to note the frequency

response of a clamp circuit. This is obtained by replacing the complex frequency s by $j\omega$. Doing so and applying simple trigonometric transformations, there results the expression

$$G_h(j\omega) = T \left[\frac{\sin \omega T/2}{\omega T/2} \right] e^{-j\omega T/2} \quad (12)$$

The amplitude and phase response of the clamp circuit are shown in Fig. 6. Effectively the response is that of a low-pass filter as had been predicted from prior discussions. On the other hand, the clamp circuit is not a perfect low-pass filter because it attenuates in the passband and admits spurious higher sampling frequencies. The latter combine to form the ripple in the reconstructed signal. The phase lag which contributes to instability is seen to reach a value of π radians at frequencies equal to the sampling frequency. Stated differently, this phase lag represents a time lag equal to one-half the sampling interval.

It should be pointed out here that more sophisticated methods of reconstructing the signal from a sequence of samples exist (Refs. 13 and 27). In general, such data holds require a considerable number of past samples to accomplish the reconstruction with a resultant tendency to cause oscillation in low-pass feedback control systems employing them. It is emphasized here that the main purpose of the data hold in a feedback control system is to improve the duty cycle of the controlled system rather than the accurate reproduction of a signal. The latter problem is most important in open-cycle communications systems where recovery of a signal regardless of the time delay is the major problem.

In feedback systems where the sampler and data hold are included inside the control loop, time lags resulting from more sophisticated reconstruction of the signal from sampled data generally result in more difficulty than is warranted. As a result, it is an almost universal custom to consider the clamp circuit a sufficient, though not perfect, data hold system. In systems employing a digital computer as a controller, it would be difficult to attempt to include a system which does more than include the contents of a register as the extrapolated sampled time function.

The remaining elements of the typical sampled-data control system shown in Fig. 1 are the controlled plant which is an analog device by its very nature. While not always linear, it is generally considered so in order to expedite analysis and synthesis. Generally, for those cases where the plant is nonlinear, it can be studied by means of linear perturbation models. In the body of theory available at the present time, the plant G is considered a linear system described fully by its transfer function expressed as a Laplace transform.

The other element shown in Fig. 1 is the error element. While shown here as a simple device which takes the difference between a reference input and a controlled output, it may be a relatively complex computer, digital or analog, which computes a control error. For perturbation models, there exists a simple linear relation between the input and output which is symbolized in the figure. In order to analyze the performance of systems containing a sampling operation, it is necessary to have available the same techniques as those which are available for linear continuous feedback systems. Subsequent sections deal with the convolution summation and the transfer function of the linear systems included in the control loop,

the combinatorial theorems which make possible the statement of overall response, condition of stability, and the requirements in a digital controller, D .

3. THE PULSE TRANSFER FUNCTION

One of the most powerful relationships which can be derived for a linear system is the one relating the output and input time functions. The concept is fairly common by now in the case of continuous systems and the transfer function which is normally used to characterize such systems is the Laplace transform of their impulsive response. Less common is the corresponding relation in sampled systems where an output pulse sequence is related to the input pulse sequence through a function known as the "pulse (or pulsed) transfer function" (Refs. 2 and 5). Having this relationship in a compact form makes the problem of analysis of sampled-data systems no more difficult or complex than for continuous systems.

To derive the pulse transfer function reference is made to Fig. 7. Here a linear system whose continuous transfer function is $G(s)$ is subjected to a sequence of pulses $r^*(t)$ derived by sampling the input function $r(t)$. The output $c(t)$ is then sampled synchronously to produce an output pulse sequence $c^*(t)$. The transforms corresponding to these quantities are $R(s)$, $R^*(s)$, $C(s)$ and $C^*(s)$ respectively.

To obtain the desired result, a convolution process is used which is clarified by reference to Fig. 8. Each pulse of the input sequence is assumed to be short enough so that an acceptable approximation to the response of $G(s)$ to this pulse is taken as the response to an impulse whose area is equal to the magnitude of the corresponding finite

pulse. Thus, at any sampling instant mT , the value of the output time function $c(t)$ is

$$c(mT) = \sum_{n=0}^{\infty} g[(m-n)T] r(nT) \quad (13)$$

where $g(m-n)T$ is the impulsive response after an elapsed time $(m-n)T$ and $r(nT)$ is the n 'th input pulse. The lower limit of n is taken as zero since it is assumed that $r(t)$ is applied at zero time.

The output pulse train $c^*(t)$ can then be expressed as in Eq. (5) by

$$c^*(t) = \sum_{m=0}^{\infty} c(mT) \delta(t - mT). \quad (14)$$

Taking the Laplace transform of this impulse sequence there results

$$C^*(s) = \sum_{m=0}^{\infty} c(mT) e^{-mTs}. \quad (15)$$

Substituting for $c(mT)$ the summation in Eq. (13),

$$C^*(s) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} g[(m-n)T] r(nT) e^{-mTs} \quad (16)$$

making a change of variable by introducing $k = (m-n)$, Eq. (15) can be rewritten

$$C^*(s) = \sum_{k=-\infty}^{\infty} \sum_{n=0}^{\infty} g(kT) r(nT) e^{-kTs} e^{-nTs}. \quad (17)$$

It is noted that since $g(kT)$ is zero for negative arguments due to the requirement for physical realizability, the lower limit of

k can be changed from $-\infty$ to zero. Hence Eq. (16) can be rearranged to

$$C^*(s) = \left[\sum_{k=0}^{\infty} g(kT) e^{-kTs} \right] \left[\sum_{n=0}^{\infty} r(nT) e^{-nTs} \right]. \quad (18)$$

The important point is that the second summation is recognized to be like that of Eq. (6) and is the pulse sequence $r^*(t)$. The first summation is the Laplace (or z -) transform of the impulsive response of the linear system. The expression in Eq. (18) can now be rewritten as

$$C^*(s) = G^*(s) R^*(s) \quad (19)$$

or

$$C(z) = G(z) R(z) \quad (20)$$

where $G^*(s)$ or $G(z)$ is the pulse transfer function.

Example: Consider the linear system whose continuous transfer function $G(s)$ is

$$G(s) = \frac{1}{s+a} \quad (21)$$

and whose impulsive response is

$$g(t) = e^{-at}. \quad (22)$$

The pulse transfer function is then

$$G(z) = \sum_{n=0}^{\infty} e^{-naT} z^{-n}. \quad (23)$$

The infinite summation so obtained can be expressed in closed form

$$G(z) = \frac{1}{1 - e^{-aT} z^{-1}} \quad (24)$$

The result could have been obtained directly from Table 1 just as for ordinary Laplace transforms.

4. INVERSION OF Z-TRANSFORMS

It is desirable to be able to invert z-transforms readily to obtain the corresponding pulse sequence. A general expression for the n'th pulse of the sequence $r^*(t)$ has been derived (Refs. 6, 11, and 12) and is given by

$$r(nT) = \frac{1}{2\pi j} \int_{\Gamma} R(z) z^{n-1} dz \quad (25)$$

where Γ is a path of integration in the z-plane that encloses all the singular points of $R(z)$. This path is the unit circle for stable systems.

For practical purposes, however, the expansion of $R(z)$ into a power series in z^{-1} by a simple process of long division (Refs. 5 and 15) can be used for inversion. This procedure is particularly useful for computing transients since it yields the initial values of the pulse sequence by the application of a simple numerical procedure. The process is best illustrated by an example.

Example: The system under consideration is shown in Fig. 9. The linear system consists of a clamp circuit and an integrator

whose combined continuous transfer function is

$$G(s) = \left[\frac{1 - e^{-Ts}}{s} \right] \left[\frac{1}{s} \right] \quad (26)$$

Use of Table 1 gives the pulse transfer function corresponding to Eq. (26)

$$G(z) = (1 - z^{-1}) \left[\frac{Tz^{-1}}{(1 - z^{-1})^2} \right] \quad (27)$$

If the input $r(t)$ is a unit step, the z-transform of the input is from Eq. (9) or Table 1,

$$R(z) = \frac{1}{1 - z^{-1}} \quad (28)$$

The output z-transform is thus

$$\begin{aligned} C(z) &= G(z) R(z) \\ &= \frac{Tz^{-1}}{1 - 2z^{-1} + z^{-2}} \end{aligned} \quad (29)$$

The inversion of $C(z)$ yields the output sequence, applying a process of long division to Eq. (29),

$$\begin{array}{r} 1 - 2z^{-1} + z^{-2} \overline{) \begin{array}{l} z^{-1} + 2z^{-2} + 2z^{-3} + \dots \\ \underline{z^{-1} - 2z^{-2} + z^{-3}} \\ 2z^{-2} - 2z^{-3} \\ \underline{2z^{-2} - 4z^{-3} + 2z^{-4}} \\ 2z^{-3} - 2z^{-4} \\ \underline{2z^{-3} - 4z^{-4} + 2z^{-5}} \\ 2z^{-4} \dots \text{etc.} \end{array}} \end{array}$$

There results a power series in z^{-1} :

$$C(z) = T [z^{-1} + 2z^{-2} + 2z^{-3} + \dots] \quad (30)$$

Each coefficient of the sequence in Eq. (30) is recognized as the magnitude of the output sample at a sample time corresponding to the order of z^{-1} . It is seen that the output is the sequence of samples resulting from sampling a ramp time function which is as it should be. The input and output pulse sequences are plotted in Fig. 10.

The example illustrates how the first group of values of a pulse sequence can be obtained readily by a simple process of long division. This is a numerical procedure which can be carried out by human computers using hand calculating machines or by means of large scale digital computers which have been programmed to implement the process of long division. On the other hand, where a general expression for the n 'th pulse is desired, the use of the contour integral of Eq. (25) can be evaluated by the method of residues. It should be pointed out also that tables such as Table 1 can be used for inversion if the z -transform is listed. If the transform is not listed, it is permissible to decompose the complex transform into partial fractions and to evaluate each fraction separately and add the results.

5. COMBINATORIAL THEOREMS

The systems which are of general interest in the design of sampled-data control systems generally have more complex configurations than the simple cascading of elements. The more complex relationships which result for feedback systems of various types can be derived by using a few simple basic relationships. These will now be outlined:

a. The pulse transfer function of two cascaded elements which are separated by a synchronous switch is the product of the

pulse transfer functions of each of the elements. Referring to Fig. 11, it is seen that the pulse sequence $C_1(z)$ is related to $R(z)$ by $G_1(z)$:

$$C_1(z) = G_1(z) R(z) \quad (31)$$

Also, the pulse sequence $C_2(z)$ is related to the pulse sequence $C_1(z)$ at the input to the second element by

$$C_2(z) = G_2(z) C_1(z) \quad (32)$$

Thus, the overall relationship between the output and input sequences is

$$C_2(z) = [G_1(z) G_2(z)] R(z) \quad (33)$$

b. The pulse transfer function of two cascaded elements not separated by a switch is the z -transform corresponding to the product of the continuous transforms. It is noted that the pulse transfer function of such a combination is not the product of the two individual pulse transfer functions. Thus, referring to Fig. 12, and recalling that the pulse transfer function is the transform of the impulsive response of the linear system, some reflection will show that the z -transform of the cascaded pair is given by

$$C_2(z) = [G_1 G_2(z)] R(z) \quad (34)$$

where $G_1 G_2(z)$ indicates the pulse transform corresponding to the system whose continuous transfer function is $G_1(s) G_2(s)$. It is emphasized once again that $G_1(z) G_2(z)$ is not equal to $G_1 G_2(z)$ as can readily be verified by trying a few examples.

By applying the basic theorems given above, it is possible to obtain overall relations between the input and output of feedback

control systems. For instance, taking the error-sampled system given in Fig. 13, the relation between input $R(z)$ and output $C(z)$ can readily be shown to be

$$C(z) = \frac{G(z)}{1 + GH(z)} R(z) \quad (35)$$

where $C(z)$ and $R(z)$ are the z -transforms corresponding to the output and input sequences, and $G(z)$ and $GH(z)$ are the pulse transforms corresponding to the feedforward transfer function $G(s)$ and loop transfer function $G(s)H(s)$.

In a similar manner, the overall relation between input and output of the system shown in Fig. 1 containing a digital stabilizer is

$$C(z) = \frac{D(z)G(z)}{1 + D(z)G(z)} R(z) \quad (36)$$

where $D(z)$ is the pulse transfer function of the digital stabilizer whose detailed significance will be shown later.

There are many other possible configurations of sampled-data systems possible besides the error sampled systems described above. The overall pulse transfer functions for a number of such systems are tabulated in Table 2. Generally speaking, the design of feedback control systems having such configurations involves the choice of elements which result in stable systems whose dynamical performance fulfills specifications set by the requirements of the system. The conditions of stability and the techniques for achieving acceptable dynamical performance will be discussed in subsequent sections.

6. STABILITY OF FEEDBACK CONTROL SYSTEMS

Of primary interest in feedback control systems is a simple method for ascertaining the stability of the system. In linear continuous systems, it can be stated that if the transfer function of the system has poles with negative real parts or stated differently, are located on the left half of the complex frequency plane, the system is stable. This statement is also applicable to sampled-data systems except that the transfer function of a sampled-data system is the ratio of polynomials in e^{Ts} rather than s and the number of poles to be surveyed is infinite. Thus, such tests as the Routh-Hurwitz criterion cannot be applied directly without some complication. The Nyquist criterion can be applied with slight modification and is quite useful in the field of sampled-data systems.

The imaginary axis of the s -plane maps into the unit circle in the e^{Ts} - or z -plane. Therefore, for roots on the s -plane having negative real parts, the magnitude in the e^{Ts} - or z -plane is less than unity. Similarly, for those roots having positive real parts, the magnitude in the e^{Ts} - or z -plane is greater than unity.

In general, therefore, the condition for stability of a linear sampled-data system is that the poles of the pulse transfer function relating input and output must lie inside the unit circle of the z -plane. Stated differently, the roots of the denominator polynomial of the pulse transfer function must have magnitudes which are less than unity. As in the case of the Nyquist criterion for linear continuous systems, the application of one of the Cauchy theorems can be used to ascertain the lack of poles outside the unit circle to prove the linear sampled system stable.

The theorem is applied by mapping a contour shown in Fig. 14 in the $G(z)$ -plane where $G(z)$ is the pulse transfer function of the system under scrutiny. Any poles of $G(z)$ which are outside the unit circle will be located in the shaded area and are enclosed by the contour. The Cauchy theorem which is used states that the map of this contour on the $G(z)$ -plane will enclose the origin of the $G(z)$ -plane a number of times equal to the difference between the zeros and poles of $G(z)$ so enclosed.

The types of configurations of interest are feedback systems whose overall transfer functions are given in Table 2. These expressions have an overall transfer function very similar to those for continuous feedback systems in that the singularities of interest stem from characteristic equations of the form $1 + GH(z) = 0$. As in the case of continuous systems, the procedure to determine stability is to ascertain the number of times the map of $GH(z)$ encloses the critical point $-1,0$ in the $G(z)$ plane.

The map of the contour shown in Fig. 14 on the $GH(z)$ plane is referred to as the pulse transfer locus and has the same status in sampled-data systems as does the Nyquist plot for continuous systems in that the enclosure of the critical points $-1,0$ is equal to the difference between the number of zeros and poles in the region outside the unit circle. If the number of poles of $GH(z)$ outside the unit circle is either zero or some known finite number, the number of zeros of $1 + GH(z)$ which are so located can be ascertained. Generally speaking, in dealing with elements which are themselves stable, the stability of the feedback control system is measured by the non-enclosure of the critical point $-1,0$.

To illustrate the point, a transfer locus for a simple sampled-data feedback control system is shown in Fig. 15. The block

diagram of the system is shown in this figure and it is seen that the pulse transfer locus does not enclose the critical point $-1,0$. Since $GH(z)$ itself contains no poles outside the unit circle it is concluded that the overall system is stable. It is further noted that if the feedforward gain were raised by a factor of slightly over two, the system would become unstable since the transfer locus would be enlarged to enclose the critical point.

Various papers have considered the problem of shaping the pulse transfer locus by adding continuous networks $N(s)$ to the continuous element $G(s)$ (Refs 4, 11, and 16). In all cases, the procedure becomes one of trial and error due to the unfortunate circumstance that the pulse transfer function of the combined elements $N(s)G(s)$ is not equal to the product of the two pulse transfer functions $N(z)$ and $G(z)$. This means that the insertion of a compensating network cannot be assessed until the entire new pulse transfer function is recomputed and replotted. Nevertheless, despite this difficulty, the stabilization of sampled-data systems by adding continuous compensating networks in the feedforward element generally follows along the lines and concepts developed for continuous systems.

By far the more important method of stabilization which is applicable to sampled-data systems is that employing a digital computer which implements a recursion formula chosen for its capacity not only to stabilize the system but also to shape its response in the time domain. It is pointed out here that even though systems might otherwise be continuous, it may pay to achieve precise transient and steady-state performance to convert them into sampled-data digitally-stabilized systems. The next section deals with the design of the system employing a digital computer to stabilize and shape a sampled-data system.

7. DESIGN OF DIGITALLY-STABILIZED FEEDBACK CONTROL SYSTEMS

As mentioned previously, one of the most important applications of sampled-data concepts is in the analysis and synthesis of systems which include a digital or an intermittent analog computer for compensation (Refs. 2, 17, 18, 21, and 22). Such a computer is capable of accepting a sequence of numbers and operating on them in some linear or nonlinear manner. By implementing a recursion relationship between the output and input sequences, it is possible to stabilize and compensate linear systems.

While the number of possible configurations is large, the one which will be considered has been shown in Fig. 1 and is reproduced in somewhat modified form in Fig. 16 for purposes of discussion. The structure is for an error-sampled and error-compensated feedback control system. For purposes of analysis the plant or controlled element is combined with the data hold preceding it and is characterized by an overall continuous transfer function $G(s)$. The digital controller has a pulse transfer function $D(z)$ which related the number sequence E_2^* at its output to the number sequence E_1^* at its input. This relationship needs some explanation as to its significance.

It is assumed that there exists a linear recursion relationship between the input and output sequences given by

$$\begin{aligned} e_2[nT] + b_1 e_2[(n-1)T] + b_2 e_2[(n-2)T] + \dots \\ = a_0 e_1[nT] + a_1 e_1[(n-1)T] + a_2 e_1[(n-2)T] + \dots \end{aligned} \quad (37)$$

where $e_2(nT)$ is the output sample at the n 'th instant etc., and $e_1(nT)$ is the output sample at the n 'th instant etc., and the coefficients a_0 , a_1 , etc., and b_0 , b_1 etc., are constants. The recursion relation given

in Eq. (37) makes possible the solution of the n 'th output sample $e_2(nT)$ by a simple linear combination of weighted previous samples at the input and output.

To obtain the pulse transfer function which implements this relationship and using the impulse approximation for the sampling process, the summation of the Laplace transforms of a sequence of impulses whose area is equal to the various e_1 's and e_2 's over all instants n is taken:

$$\begin{aligned} \sum_{n=0}^{\infty} e_2(nT) z^{-n} + b_1 \sum_{n=0}^{\infty} e_2(nT) z^{-n-1} + b_2 \sum_{n=0}^{\infty} e_2(nT) z^{-n-2} \\ = a_0 \sum_{n=0}^{\infty} e_1(nT) z^{-n} + a_1 \sum_{n=0}^{\infty} e_1(nT) z^{-n-1} \\ + a_2 \sum_{n=0}^{\infty} e_1(nT) z^{-n-2} \end{aligned} \quad (38)$$

where the increased negative orders of z are used for the purpose of shifting the sequences the necessary integral number of sample times to take into account the delayed pulses indicated in the recursion formula.

It is possible to factor out the summations as shown in the following:

$$\begin{aligned} \sum_{n=0}^{\infty} e_2(nT) z^{-n} \{1 + b_1 z^{-1} + b_2 z^{-2} + \dots\} \\ = \sum_{n=0}^{\infty} e_1(nT) z^{-n} \{a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots\} \end{aligned} \quad (39)$$

The summations which have been factored out in Eq. (39) are recognized to be the z -transforms of the output and input pulse sequences, $E_2(z)$ and $E_1(z)$ respectively.

Replacing the infinite summations by these equivalents and rearranging,

$$\begin{aligned} E_2(z) \{1 + b_1 z^{-1} + b_2 z^{-2} + \dots\} \\ = E_1(z) \{a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots\} \end{aligned} \quad (40)$$

which is expressed as a pulse transfer function defined as follows:

$$D(z) = \frac{E_2(z)}{E_1(z)} = \frac{a_0 + a_1 z^{-1} + a_2 z^{-2} + \dots}{1 + b_1 z^{-1} + b_2 z^{-2} + \dots} \quad (41)$$

The pulse transfer function of a linear digital process so expressed is conveniently handled in analysis and synthesis. The linear recursion formula which it represents can readily be implemented by an intermittent computer whether it be of the digital or analog form or some convenient combination. The problem of synthesis resolves itself to specifying the various constants, a_0, a_1, a_2, \dots and b_1, b_2, \dots which are required to achieve some overall performance specification.

The overall response of the system is expressed by the overall pulse transfer function $C(z)/R(z)$ which is written as $K(z)$. Desirable forms of $K(z)$ are known as prototypes and can be specified on the basis of requirements of the system performance, such as the ability to settle in a specified time, the ability to follow perfectly certain specified inputs in the steady state and time domain, and other specifications. The important point is that once an overall pulse transfer function $K(z)$ is specified, it is possible to achieve it by means of a digital stabilizer, subject to certain limitations which will be discussed later.

To demonstrate how this is done, consider the overall response of the system shown in Fig. 16,

$$K(z) = \frac{C(z)}{R(z)} = \frac{D(z) G(z)}{1 + D(z) G(z)} \quad (42)$$

Since $D(z)$ is adjustable and under the control of the designer Eq. (43) is solved for $D(z)$ in terms of $K(z)$:

$$D(z) = \frac{1}{G(z)} \frac{K(z)}{1 - K(z)} \quad (43)$$

The relatively simple process outlined above is subject to certain limitations which will be given later. Before going into these limitations, it is worthwhile to discuss now the forms of the desired overall pulse transfer functions $K(z)$ which are to be achieved.

The overall response $K(z)$ determines the manner in which the output of a feedback system is affected by the input. For instance, in a duplicator control system, the desired response is that the output reproduce the input at all times. Transients which occur should disappear as quickly as possible or should fulfill set specifications of rise time and overshoot. On the other hand, in the case of a regulator where the inputs are undesirable disturbances, the output should be completely insensitive to the disturbance inputs. Any transients should likewise be minimum or fulfill set specifications. Thus, $K(z)$ contains the overall requirements of the control system and it can be realized by introducing a digital controller.

An important class of overall response functions is the one which can be referred to as having finite settling time. In a system of this type, the transient decays in a finite time. To clarify the point, it should be noted that the transient referred to is the sampled

output so that settling takes place at the sample times only. Between sampling instants, the output will contain ripple, which in most systems will also disappear gradually. Finite settling time systems are characterized by response functions $K(z)$ which are polynomials. For instance, the very simplest function is

$$K(z) = z^{-1}. \quad (44)$$

For a function of this type, the response settles to its steady-state value after one sample interval. Actually, as a general rule, the continuous output $c(t)$ oscillates about the steady-state value but crosses the steady-state value at sampling instants. The term "finite settling time" used in these discussions refers to the output as reached at sample instants only rather than the continuous output itself.

The requirement that response functions of duplicator systems follow such inputs as step functions, ramp or constant velocity input functions, and parabolic or constant acceleration functions can be met by considering the control error E_1 in Fig. 16. It can readily be shown the $E_1(z)$ is given by

$$E_1(z) = R(z)[1 - K(z)]. \quad (45)$$

If the system follows the input perfectly in the steady state, the control error goes to zero. Thus applying the final value theorem for z -transforms (Ref. 11), the final value of the control error E_{1ss} is given by

$$E_{1ss} = \left[(1 - z^{-1}) E_1(z) \right]_{z \rightarrow 1}. \quad (46)$$

Now for the case where the input is a step, ramp, or parabola,

$$R(z) = \frac{M(z)}{(1 - z^{-1})^n} \quad (47)$$

where $M(z)$ is a polynomial in z^{-1} and n is the degree of the input ($n = 1$ for step, 2 for a ramp etc.). Thus, the steady-state error is

$$E_{1ss} = \frac{[1 - K(z)]}{(1 - z^{-1})^{n-1}} M(z). \quad (48)$$

Now in order for the steady-state error in following to be zero, it is necessary that $[1 - K(z)]$ contain the factor $(1 - z^{-1})$ to the one higher degree as that of the denominator of the input function or equal in degree to the denominator of the z -transform of the input, $R(z)$.

For convenience, a tabulation of minimal prototypes which satisfy the requirements above are given in Table 3 (Ref. 17). These response functions are called minimal because they represent the simplest functions that can be used which will satisfy the requirement. It is understood that much more complex functions can be devised which will satisfy the requirement since the latter states only that the term $(1 - z^{-1})$ must be a factor of $[1 - K(z)]$ in the correct degree but that any other factors may be contained if desired. It is thus possible to specify overall transfer functions having finite settling time and the capacity for following steps, ramps, or other polynomial inputs in the steady state.

Generally speaking, systems having finite settling time and conditions which permit the perfect steady state following of a given higher order polynomial such as a constant acceleration input will exhibit poor transient performance when subjected to a polynomial of lower order such as a step function. For systems which are subjected to a variety of inputs, the use of a "staleness factor" in the overall transfer function is recommended.

The staleness factor introduces a smoothing effect in the overall response very similar to that produced in a continuous system by a low-pass resistance-capacitance network. The term was originally introduced by Barker (Ref. 2) and extended by the work of Bertram (Ref. 19). The prototype overall transfer function must satisfy all the requirements outlined previously for the following of a polynomial but also contain a denominator as shown below

$$K(z) = \frac{(1 - z^{-1})^n F(z)}{(1 - az^{-1})^m} \quad (49)$$

where a is known as the "staleness factor." A little thought will show that a response of this type does not have a finite settling time even at sampling instants but will approach a steady-state condition at infinite time as is the case in all continuous linear systems.

Bertram shows that the constant a should have values ranging from 0.4 to 0.6 and that the order of the staleness factor term in the denominator of $K(z)$ need be not higher than the first although Barker recommends higher orders. In any case, the compromise between desirable step, ramp, and constant acceleration response functions can be achieved by the introduction of staleness factors as indicated. It is emphasized that the overall response functions can be achieved by the introduction of a digital stabilizer element and the subsequent material indicates how the transfer functions for such devices are obtained.

The basic specification of the digital control transfer function is contained in Eq. (44). Once the overall transfer function $K(z)$ is specified and the plant transfer function $G(z)$ is known, the required digital stabilizer transfer function is also specified. There are however, certain limitations in

the free choice of $K(z)$. After satisfying the requirements for following the polynomial inputs in the steady state, the additional requirements which $K(z)$ must satisfy are given as follows (Refs. 17 and 19).

a. The overall pulse transfer function $K(z)$ must contain as zeros all those zeros of $G(z)$ which are on or outside the unit circle in the z -plane.

b. The function $[1 - K(z)]$ must contain as zeros all those poles of $G(z)$ which lie on or outside the unit circle in the z -plane.

Once these limitations have been met, the digital program for stabilization and shaping, $D(z)$, can be specified. The design of such a controller can best be illustrated by means of an example.

Example: The system to be designed is shown in Fig. 17 where it is seen that the plant to be controlled is a single integrator with two equal time delays. The plant is preceded by a clamp hold circuit and the error is sampled and processed in a digital stabilizer marked $D(z)$. It can be readily ascertained that even if the system were continuous with the sample and hold operation removed and replaced with a direct connection, the system would be unstable. Introduction of sampling makes the system even more unstable than the continuous system.

It is desired to design the digital program represented by $D(z)$ which will make the system stable, have a finite settling time and be capable of following a constant velocity (ramp) input perfectly in the steady state. The pulsed transfer function $G(z)$ of the plant and data hold combined is given by

$$G(z) = \mathcal{Z} \left[\frac{1 - e^{-Ts}}{s^2(s-1)} \right] \quad (50)$$

This transform can be found in the complete tables given by Barker (Ref. 2) or by expansion of $G(s)$ into partial fractions and obtaining the pulse transfer functions from the more restricted Table 1. This pulse transfer function is

$$G(z) = \frac{(1 - 2.34z^{-1})(1 + 0.16z^{-1})z^{-1}}{(1 - z^{-1})(1 - 0.368z^{-1})} \quad (51)$$

Examining the pulse transfer function, it is seen that there is a zero in $G(z)$ which lies at 2.34 or outside the unit circle. Hence to meet the limitation set forth previously, it is necessary that the overall transfer function $K(z)$ contain this zero as one of its zeros. Thus, $K(z)$ must be

$$K(z) = (1 - 2.34z^{-1})(a_1z^{-1} + a_2z^{-2}) \quad (52)$$

where a_1 and a_2 are unspecified coefficients.

By making $K(z)$ a polynomial, a finite settling time is assured. In order to meet the specification that the system follow a ramp function perfectly in the steady state, $[1 - K(z)]$ must contain $(1 - z^{-1})$ to the second order. This condition results in the requirement that

$$[1 - K(z)] = (1 - z^{-1})(1 + bz^{-1}) \quad (53)$$

where b_1 is an unspecified coefficient.

The overall pulse transfer function $K(z)$ must satisfy simultaneously the conditions imposed by Eqs. (52) and (53). Substituting $K(z)$ from Eq. (52) in Eq. (53), there results the equality

$$\begin{aligned} 1 - a_1z^{-1} - (2.34a_1 + a_2)z^{-2} - 2.34a_2z^{-3} \\ = 1 + (b_1 - 2)z^{-1} + (1 - 2b_1)z^{-2} + b_1z^{-3} \end{aligned} \quad (54)$$

which can be satisfied if the coefficients of terms of like order are equal, thus resulting in the following simultaneous conditions:

$$\begin{aligned} b_1 - 2 &= -a_1 \\ 2b_1 - 1 &= 2.34a_1 + a_2 \\ b_1 &= -2.34a_2 \end{aligned} \quad (55)$$

There being three equations and three unknowns, a solution is obtained giving values for the coefficients of $K(z)$ as follows:

$$\begin{aligned} a_1 &= 0.81 \\ a_2 &= -0.51 \end{aligned} \quad (56)$$

It is noted that the minimum number of arbitrary coefficients were chosen to satisfy the conditions of the problem. If more terms in higher orders of z^{-1} were used in $K(z)$ resulting in a longer settling time, there would have been more constants than conditions and all coefficients except two could be arbitrarily specified. This means that transient response could be controlled at the expense of more terms in $K(z)$. As solved here, the system is a minimal one since the settling time is at a minimum.

With the coefficients of $K(z)$ used as given in Eq. (56), the overall response of the system is

$$K(z) = 0.81z^{-1} + 1.38z^{-2} - 1.19z^{-3} \quad (57)$$

If the input function $R(z)$ is a unit constant velocity (ramp) the output of the system $C(z)$ is given by

$$C(z) = \frac{(0.81z^{-1} + 1.38z^{-2} - 1.19z^{-3})z^{-1}}{(1 - z^{-1})^2} \quad (58)$$

The pulse sequence corresponding to this result is plotted in Fig. 18 where it is seen that the system produces a zero following error at sample instants after three sample intervals. The command to the plant $E_2(z)$ is also plotted where it is seen that an offset value is produced at steady state. It is this offset which makes it possible for the system to follow a ramp input perfectly even though the plant has only a single integration which would normally cause the following error to be some finite constant.

The digital stabilizer which is required is obtained by substitution of $K(z)$ and $G(z)$ in Eq. (43). Doing this, the pulse transfer function of the digital stabilizer $D(z)$ becomes

$$D(z) = \frac{0.81 - 1.106z^{-1} + 0.485z^{-2} - 0.691z^{-3}}{1 + 0.352z^{-1} - 1.159z^{-2} - 0.193z^{-3}} \quad (59)$$

The significance of this pulse transfer function in terms of a recursion formula is given in Eq. (37) and a system diagram showing the sequence of numerical operations is shown in Fig. 19. This diagram illustrates the process of holding each sample datum, multiplying it by its appropriate weight, delaying it, and adding all the weighted numbers to obtain the present output sample $e_2(nT)$.

The example given above is an illustration of how a digital computer can be programmed to stabilize and produce desirable dynamic response characteristics in a linear system. The prototype $K(z)$ which was chosen is a minimal finite settling type in that it follows the input perfectly after a transient of finite duration. Generally speaking, systems of this type have excessive overshoots when subjected to inputs for which they are not specifically designed.

A better compromise is reached by using staleness factors as suggested by Barker and Bertram (Refs. 2 and 19). Another point of interest is that the system follows the ramp perfectly only at sample instants. Between these instants, the system oscillates in and out of the correct position at sampling frequency. This effect is known as ripple and means are available to study its magnitude (Refs. 21, 22, 23, and 24). Ripple effects are generally reduced when staleness factors are used.

While the emphasis has been placed on a given prototype employing the digital stabilizer in the error line, other forms can be similarly analyzed and synthesized when the digital stabilizer is in the feedback line or with combination systems where the digital stabilizer bypasses a continuous error line (Ref. 20).

Behavior of sampled-data systems in the presence of random inputs has been studied by Barker, Franklin, and Lees (Refs. 25, 26, and 27). The theory will not be presented here except to state that many of the functions commonly used in the analysis of continuous systems are found in sampled systems as well. For instance, there are defined the correlation function and the power spectrum for sampled random functions. Optimization procedures by minimization of the mean square system error are available. It may be stated generally that sampled data control systems can be optimized in much the same manner as the continuous systems, possibly with more flexibility since there is generally a digital stabilizer which can implement overall desired transfer functions as required in the optimization procedure.

Other sampled-data problems which have been studied are those in which a multiplicity of sampling operations at different sampling frequencies are present (Refs. 28 and 29).

This situation arises in systems having simultaneously a sampled-data information gathering device such as a radar; a digital computer whose output data rate may be lower, equal to, or higher than the radar data rate; and one or more intermittent digital data links. While the analysis of such systems is not simple, it is sufficiently regular to make the design of such systems practical.

8. APPLICATION OF SAMPLED-DATA SYSTEMS TO THE GUIDANCE OF MISSILES

It is the purpose of this section to point out the application of sampled-data systems in the guidance and control of missiles. Guidance systems may be roughly divided into two classes, those in which data gathering and course computation is implemented by devices carried by the missile; and those in which such devices are situated on the ground, mixed systems have combinations in which both missile-borne and ground equipment are used. Typical missile-borne systems are those employing such guidance systems as preprogramming, homing, beam-riding, and inertial and celestial navigation. Due to weight and space limitations, it is not likely that missile-borne digital data and processing devices be used to compute control commands. On the other hand, systems using ground-based instrumentation often have sampled elements such as data-gathering radars, data-processing computers, and ground-to-air data links. For this reason, the expected incidence of sampled-data systems is highest in radio command guidance systems.

One of the advantages of ground-based systems is that they provide a known and stable ground reference on which commands can be based. Also, the possibility of using larger computers makes practical the implementation of more sophisticated data-processing, guidance, and decision making

procedures. Nevertheless, even though space and weight are not major limitations, the computer should be as simple as possible in the interests of economy and reliability. Therefore, unnecessarily high-speed computers which would result in a high output data rate are to be avoided and the lowest data rates consistent with satisfactory overall performance of the missile system should be used. It is precisely under such conditions of low data or sampling rates that sampled data theory will provide a sound assessment of performance, stability, and accuracy of guidance systems.

As an illustration of the application of sampled-data techniques, consider a hypothetical system employing a digital computer for data reduction and for the computation of steering commands. The system is assumed to be a ground-to-air defensive missile system employing a radio command guidance system. It is assumed that the ground-based radar data gathering systems are pencil beam tracking radars disposed on the ground to form a triangulation system for locating both defensive and offensive missiles. The primary data are range from each radar site to both missiles so that the digital computer must reduce these data to obtain fixes relative to a reference set of ground coordinates. A missile track is established and the future position of the offensive missile predicted by the computer. A midcourse track for the defensive missile is computed and stored in the computer.

The object of the midcourse guidance is to steer the defensive missile to follow the track as accurately as possible. Deviations of the missile from this desired track are converted into a control error periodically and a steering command is transmitted to the missile every cycle time of the computer. The steering commands are computed by a digital program $D(z)$ obtained by techniques described in previous sections leading to both a stable and relatively docile system.

It is noted that a steering command can be delivered by the computer only at sampling intervals determined by the cycle time required to reduce data to obtain a present position fix, implement the recursion formula which yields the steering command, and transmit this command to the missile. The time required to complete the cycle is the sampling time which, in the interest of economy, should be as large as possible consistent with response requirements of the missile. This system is a sampled-data system and must be designed as such.

It should be pointed out that not all sampled-data systems are of the magnitude or complexity of the one described in the hypothetical example given above. For instance, a digital time-shared data link can transmit a command to a missile only periodically, thus interposing a sampling operation in the control loop. It should also be pointed out that digital controllers need not employ complex large scale digital computers. Simple and compact special purpose sampled controllers can be developed which take advantage of the relatively slow data rate that is required in most systems and which use optimum combinations of digital and analog techniques. However, where the computer is required for the purpose of reducing data as described above, it is just as well to program the control function as part of the cycle.

As seen from the brief discussion above, sampled-data control systems are a natural extension of the concept of ground command guidance employing digital computers. On the other hand, with the accelerated development in the design and packaging of digital devices, it is not beyond the realm of imagination to conceive of missile-borne digital devices for the processing of data obtained directly from instrumentation carried by the missile and to generate periodic steering commands. For instance, an inertial

system requiring the double integration of the output of accelerometers to obtain position can be carried out by numerical methods employing optimum quadrature formulas. This operation can be implemented by means of a digital computer as can other data processing operations and, in addition, compute steering commands which will stabilize and control the missile on its course. If this be the case, even airborne systems will require analysis and synthesis by means of sampled-data theory.

The employment of digital devices for guidance and control permits the use of increased sophistication in control. The possibility of designing adaptive systems which adjust the program for computing the steering command in accordance with external conditions is by no means remote. It is now possible to design linear time variant systems which are programmed to change their characteristics with time (Refs. 30 and 31). Such systems can be made economical in expenditure of control effort and terminate the missile flight accurately. In addition, computers can be programmed to implement simple logical decisions required before and during the flight of the missile. It is because of the flexibility of the digital computer that such advanced systems can be implemented either by use of a ground-based computer or, less readily, by means of a missile-borne computer.

9. CONCLUSIONS

The sampled-data control system has a place in the general area of guidance and control of missiles. The theory which has been presented is only a small part of the body of theory which is now available and which is in the process of development. Because of the fact that sampled-data systems often go hand in hand with digital computers,

it is likely that a higher degree of sophistication can be achieved with such systems. On the other hand, the designer is often faced with the necessity of designing a system as a sampled-data theory because one or more of the elements, such as data gathering radars and data links, is intermittent in nature.

The sampled-data system has a close relationship to sampled systems which are studied in the theory communications. As in the latter, some information is lost or deteriorated by the process of sampling as is indicated by a simple application of the sampling theorem. Yet, in most cases, a completely continuous system has the capability of transmitting a far greater

amount of information than is really required. The choice of sampling rates can be arrived at rationally unless dictated by other considerations. Having settled on this figure, rational design procedures are available for the design of linear time variant and time-invariant control and transmission systems. Not yet fully understood is the behavior and design of nonlinear and adaptive sampled-data systems although some efforts and results have been obtained in this direction (Ref. 32). In conclusion, the sampled-data control system has considerable present and potential importance in the field of guidance and control of missiles and an understanding of the operation of such systems should be of considerable value to the designer.

REFERENCES

1. Hurwicz, W., Chapter 5, "Filters and Servo Systems with Pulsed Data," in James, H. M., Nichols, N. B., and Phillips, R. S., "Theory of Servomechanisms," McGraw-Hill Book Co., New York, N. Y., Vol. 25, Radiation Laboratory Series, 1947.
2. Barker, R. H., "The Pulse Transfer Function and Its Application to Sampling Servo Systems," Proc. I. E. E., (London), Part IV, Monograph No. 43, July 15, 1952.
3. Barker, R. H., "The Theory of Pulse-Monitored Servos and Their Use for Prediction," Report No. 1046, Signals Research and Development Establishment, Christchurch, Hantz, England, November 1950.
4. Linvill, W. K., "Sampled-Data Control Systems Studied Through Comparison with Amplitude Modulation," Trans. A. I. E. E., Vol. 70, Part II, 1951, Pages 1779-88.
5. Ragazzini, J. R. and Zadeh, L. A., "The Analysis of Sampled-Data Systems," Trans. A. I. E. E., Vol. 71, Part II, November 1952, pp. 225-34.
6. Lawden, D. F., "A General Theory of Sampling Servo Systems," Proc. I. E. E., (London, England), Vol. 98, Part IV, October 1951.
7. Tsipkin, Y. Z., "Theory of Intermittent Regulation," *Automatika i Telemekhanika*, (Moscow, USSR), Vol. 10, No. 3, 1949, pp. 189-224.
8. Tsipkin, Y. Z., "Frequency Method of Analyzing Intermittent Regulating Systems," *Automatika i Telemekhanika*, Vol. 14, No. 1, 1953, pp. 11-33.

9. Brown, B. M., "Application of Finite Difference Operators to Linear Systems," Proc. D. S. I. R. Conference on Automatic Control, Butterworths Scientific Publications, (London, England), Edited by A. Tustin, 1952.
10. Raymond, F. H., "Analysis of Discontinuous Servomechanisms," Annales des Telecommunications, (Paris, France), Vol. 4, pp. 250-56, July 1949, pp. 307-314; August-September 1949; pp. 347-357, October 1949.
11. Jury, E. I., "Analysis and Synthesis of Sampled-Data Systems," Trans. A. I. E. E., Part I, Vol. 73, 1954.
12. Holt Smith, C., Lawden, D. F., Bailey, A. E., "Characteristics of Sampling Servo Systems," Proc. D. S. I. R. Conference on Automatic Control, Butterworths Scientific Publications, (London, England), Edited by A. Tustin, 1952.
13. Porter, A. and Stoneman, F., "A New Approach to the Design of Pulse-Monitored Servo Systems," Proc. I. E. E., (London, England), Part II, Vol. 97, p. 597, 1950.
14. Salzer, J. M., "Treatment of Digital Control Systems and Numerical Processes in the Frequency Domain," Sc. D. Thesis, Dept. of E. E., M. I. T., Cambridge, Massachusetts, 1947.
15. Ragazzini, J. R. and Bergen, A. R., "A Mathematical Technique for the Analysis of Linear Systems," Proc. I. R. E., Vol. 42, No. 11, November 1954, pp. 1645-1651.
16. Sklansky, J., "Network Compensation of Error-Sampled Feedback Control Systems," Technical Report T-7/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., April 1, 1955.
17. Bergen, A. R. and Ragazzini, J. R., "Sampled-Data Processing Techniques for Feedback Control Systems," Trans. A. I. E. E., Vol. 73, 1954.
18. Linvill, W. K. and Salzer, J. M., "Analysis of Control Systems Involving a Digital Computer," Proc. I. R. E., Vol. 41, No. 7, pp. 901-906, 1953.
19. Bertram, J. E., "Factors in the Design of Digital Controllers for Sampled-Data Feedback Control Systems," Trans. A. I. E. E., Paper No. 56-209, 1956.
20. Maitra, K. K., and Sarachik, P. E., "Digital Compensation of Continuous-Data Feedback Control Systems," Trans. A. I. E. E., Part II, Vol. 76, No. 24, May 1956.
21. Linvill, W. K. and Sittler, R. W., "Extension of Conventional Techniques to the Design of Sampled-Data Systems," Convention Record, I. R. E., Part I, 1953, pp. 99-104.
22. Lago, G. V. and Truxal, J. G., "The Design of Sampled-Data Feedback Systems," Trans. A. I. E. E., Part II, Vol. 74, November 1954, pp. 247-252.
23. Sklansky, J. and Ragazzini, J. R., "Analysis of Errors in Sampled-Data Feedback Systems," Trans. A. I. E. E. Part II, Vol. 75, May 1955.

24. Jury, E. I., "Synthesis and Critical Study of Sampled-Data Control Systems," Trans. A. I. E. E., Part II, Paper No. 56-208, July 1956.
25. Franklin, G. F., "Linear Filtering of Sampled Data," Technical Report T-5/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., December 1954.
26. Franklin, G. F., "The Optimum Synthesis of Sampled-Data Systems," Technical Report T-6/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., May 2, 1955.
27. Lees, A. B., "Interpolation and Extrapolation of Sampled Data," Trans. I. R. E., Professional Group on Information Theory, Vol. 1T-2, No. 1, March 1956.
28. Kranc, G. M., "The Analysis of Multiple-Rate Sampled Systems," Technical Report T-11/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., September 15, 1955.
29. Kranc, G. M., "Multi-Rate Sampled Systems," Technical Report T-14/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., May 7, 1956.
30. Friedland, B., "A Technique for the Analysis of Time-Varying Sampled-Data Systems," Technical Report T-10/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., September 15, 1955.
31. Friedland, B., "Transformation Techniques for Time-Varying Sampled-Data Systems," Technical Report T-13/B, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., January 2, 1956.
32. Kalman, R. E., "Investigation of Non-Linear Control Systems Operating on Sampled Data," Technical Note TN-4/127, Electronics Research Laboratories, Department of Electrical Engineering, Columbia University, New York 27, N. Y., July 31, 1956.

ADDITIONAL BIBLIOGRAPHY

- B-1. Laplace, P. S., "Theorie Analytic des Probabilites, Part I: Du Calcul des Fonctions Generatrices," (book), Paris, France, 1812.
- B-2. deMoivre, A., "Miscellanea Analytica de Seriebus et Quadraturis," London, England, 1730.
- B-3. Seal, H. L., "The Historical Development of the Use of Generating Functions in Probability Theory," Mitteilungen der Vereinigung Schweizerischer Versicherungs Mathematiker, Bern, Switzerland, Vol. 49, pp. 209-228, 1949.
- B-4. MacColl, L. A., "Fundamental Theory of Servomechanisms," (book), D. Van Nostrand and Co., New York, N. Y., 1945, Chapter X.

- B-5. Stone, W. M., "A List of Generalized Laplace Transforms," Iowa State College Journal of Science, Vol. 22, No. 3, pp. 215-225, April 1948.
- B-6. Truxal, J. G., "Automatic Feedback Control System Synthesis," (book), McGraw-Hill Book Co., New York, N. Y., 1955, Chapter 9.
- B-7. Samuelson, P. A., "Foundations of Economic Analysis," (book), Harvard University Press, Cambridge, Massachusetts, 1947.
- B-8. Miller, K. S. and Schwarz, R. J., "Analysis of Sampled-Data Servomechanisms," Journal of Applied Physics, Vol. 21, No. 4, April 1950, pp. 290-294.
- B-9. Brown, R. G. and Murphy, G. L., "An Approximate Transfer Function for the Analysis and Design of Pulsed Servos," Trans. A. I. E. E., Vol. 71, Part II, 1952, pp. 435-440, (1953 section).
- B-10. Johnson, G. W. and Lindorff, D. P., "Transient Analysis of Sampled-Data Control Systems," Trans. A. I. E. E., Part II, Vol. 74, July 1954, pp. 147-153.
- B-11. Jury, E. I., "The Effect of Root Locations on the Transient Response of Sampled-Data Systems," Trans. A. I. E. E., Part II, Vol. 75, March 1955.
- B-12. Lago, G. V., "Additions to Z-transformation Theory for Sampled-Data Systems," Trans. A. I. E. E., Part II, Vol. 75, January 1955, pp. 403-408.
- B-13. Teichmann, T., "Closed-Loop Control System Containing a Digital Computer," Trans. I. R. E. Group on Electronic Computers, Vol. EC-4, No. 3, September 1955, pp. 106-117.
- B-14. Salzer, J. M., "Frequency Analysis of Digital Computers Operating in Real Time," Proc. I. R. E., Vol. 42, No. 2, February 1954, pp. 457-466.
- B-15. Chow, C. K., "Contactor Servomechanisms Employing Sampled-Data," Trans. A. I. E. E., Part II, Vol. 74, March 1954, pp. 51-62.
- B-16. Russell, F. A., "Design Criterion for Stability of Sampled-Data On-Off Servomechanisms," Doctoral Thesis, Department of Electrical Engineering, Columbia University, New York 27, N. Y., June 1953.
- B-17. Jury, E. I., "Discrete Compensation of Sampled-Data and Continuous Systems," Trans. A. I. E. E., Paper 56-644, 1956.
- B-18. Freeman, Herbert, "Multipole Sampled-Data Control Systems," Technical Report T-12/B, Electronics Research Laboratory, Department of Electrical Engineering, Columbia University, New York 27, N. Y., September 30, 1955.

Table 1. Abbreviated Table of Laplace and z-Transforms

	Laplace Transform $F(s)$	Time Function $f(t)$	z-Transforms $F^*(z)$
(1)	1	$\delta(t)$	z^{-0}
(2)	e^{-nTs}	$\delta(t - nT)$	z^{-n}
(3)	$\frac{1}{s}$	1	$\frac{1}{1 - z^{-1}}$
(4)	$\frac{1}{s^2}$	t	$\frac{Tz^{-1}}{(1 - z^{-1})^2}$
(5)	$\frac{1}{s + a}$	e^{-at}	$\frac{1}{(1 - e^{-aT}z^{-1})}$
(6)	$\frac{a}{s(s + a)}$	$(1 - e^{-at})$	$\frac{z^{-1}(1 - e^{-aT})}{(1 - z^{-1})(1 - e^{-aT}z^{-1})}$
(7)	$\frac{a}{s^2 + a^2}$	$\sin at$	$\frac{\sin aTz^{-1}}{1 - (2 \cos aT)z^{-1} + z^{-2}}$
(8)	$F(s + a)$	$e^{-at}f(t)$	$F^*(e^{-aT}z)$

Table 1. Abbreviated Table of Laplace and z -Transforms (Continued)

	Laplace Transform $F(s)$	Time Function $f(t)$	z -Transform $F^*(z)$
(9)	$e^{-sT} F(s)$	$f(t-T)$	$z^{-1} F^*(z)$
(10)	$\frac{1}{s - \frac{1}{T} \ln a}$	$a^{t/T}$	$\frac{z}{z - a}$
(11)	$\frac{s}{s^2 + a^2}$	$\cos at$	$\frac{1 - \cos aT z^{-1}}{1 - (2 \cos aT) z^{-1} + z^{-2}}$
(12)	$\frac{a}{s^2(s+a)}$	$\frac{1}{a} (at - 1 + e^{-at})$	$\frac{z^{-1}}{(1 - z^{-1})^2} - \frac{(1 - e^{-aT})z^{-1}}{a(1 - z^{-1})(1 - e^{-aT}z^{-1})}$
(13)	$\frac{b-a}{(s+a)(s+b)}$	$e^{-at} - e^{-bt}$	$\frac{e^{-aT} - e^{-bT}}{(1 - e^{-aT}z^{-1})(1 - e^{-bT}z^{-1})}$
(14)	$\frac{s+a}{(s+a)^2 + b^2}$	$e^{-at} \cos bt$	$\frac{1 - z^{-1}e^{-aT} \cos bT}{1 - 2z^{-1}e^{-aT} \cos bT + e^{-2aT}z^{-2}}$

Table 2. Output Transforms for Basic Sampled-Data Systems

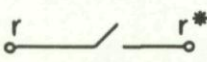
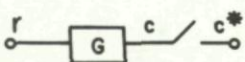
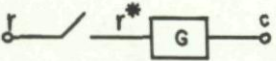
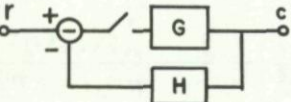
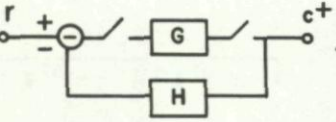
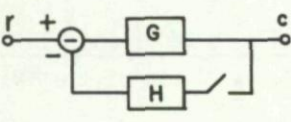
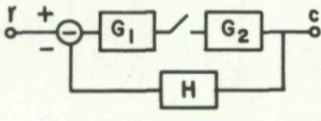
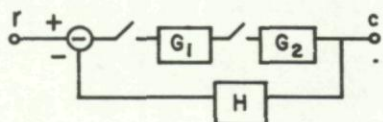
System	Laplace Transform of Output $C(s)$	z -Transform of Output $C^*(z)$
(1) 	$\dots \dots \dots R^*(s) \dots \dots \dots$	$\dots \dots \dots R^*(z) \dots \dots \dots$
(2) 	$\dots \dots \dots GR^*(s) \dots \dots \dots$	$\dots \dots \dots GR^*(z) \dots \dots \dots$
(3) 	$\dots \dots \dots G(s)R^*(s) \dots \dots \dots$	$\dots \dots \dots G^*(z)R^*(z) \dots \dots \dots$
(4) 	$\dots \dots \dots \frac{G(s)R^*(s)}{1 + HG^*(s)} \dots \dots \dots$	$\dots \dots \dots \frac{G^*(z)R^*(z)}{1 + HG^*(z)} \dots \dots \dots$
(5) 	$\dots \dots \dots \frac{G^*(s)R^*(s)}{1 + H^*(s)G^*(s)} \dots \dots \dots$	$\dots \dots \dots \frac{G^*(z)R^*(z)}{1 + H^*(z)G^*(z)} \dots \dots \dots$
(6) 	$\dots \dots \dots G(s) \left[R(s) - \frac{H(s)RG^*(s)}{1 + HG^*(s)} \right] \dots \dots \dots$	$\dots \dots \dots \frac{RG^*(s)}{1 + HG^*(z)} \dots \dots \dots$
(7) 	$\dots \dots \dots \frac{G_2(s)RG_1^*(s)}{1 + HG_1G_2^*(s)} \dots \dots \dots$	$\dots \dots \dots \frac{G_2^*(z)RG_1^*(z)}{1 + HG_1G_2^*(z)} \dots \dots \dots$
(8) 	$\dots \dots \dots \frac{G_2(s)G_1^*(s)R^*(s)}{1 + G_1^*(s)G_2H^*(s)} \dots \dots \dots$	$\dots \dots \dots \frac{G_1(z)G_2(z)R(z)}{1 + G_1(z)G_2H(z)} \dots \dots \dots$

Table 3. Transfer Functions Corresponding to Prototype Overall Transmissions

Overall Transmission $K^*(z)$	Transfer Function $D^*(z)G^*(z)$
For Step Input Test Function, $R(s) = 1/s$	
z^{-1}	$\frac{z^{-1}}{1 - z^{-1}}$
z^{-2}	$\frac{z^{-2}}{(1 - z^{-1})(1 + z^{-1})}$
z^{-3}	$\frac{z^{-3}}{(1 - z^{-1})(1 + z^{-1} + z^{-2})}$
z^{-4}	$\frac{z^{-4}}{(1 - z^{-1})(1 + z^{-1} + z^{-2} + z^{-3})}$
For Ramp Input Function, $R(s) = 1/s^2$	
$2z^{-1} - z^{-2}$	$\frac{2z^{-1} - z^{-2}}{(1 - z^{-1})^2}$
$3z^{-2} - 2z^{-3}$	$\frac{3z^{-2} - 2z^{-3}}{(1 - z^{-1})^2 (1 + 2z^{-1})}$
$4z^{-3} - 3z^{-4}$ etc.	$\frac{4z^{-3} - 3z^{-4}}{(1 - z^{-1})^2 (1 + 2z^{-1} + 3z^{-2})}$
For Acceleration Input Function, $R(s) = 1/s^3$	
$3z^{-1} - 3z^{-2} + z^{-3}$ etc.	$\frac{3z^{-1} - 3z^{-2} + z^{-3}}{(1 - z^{-2})^3}$

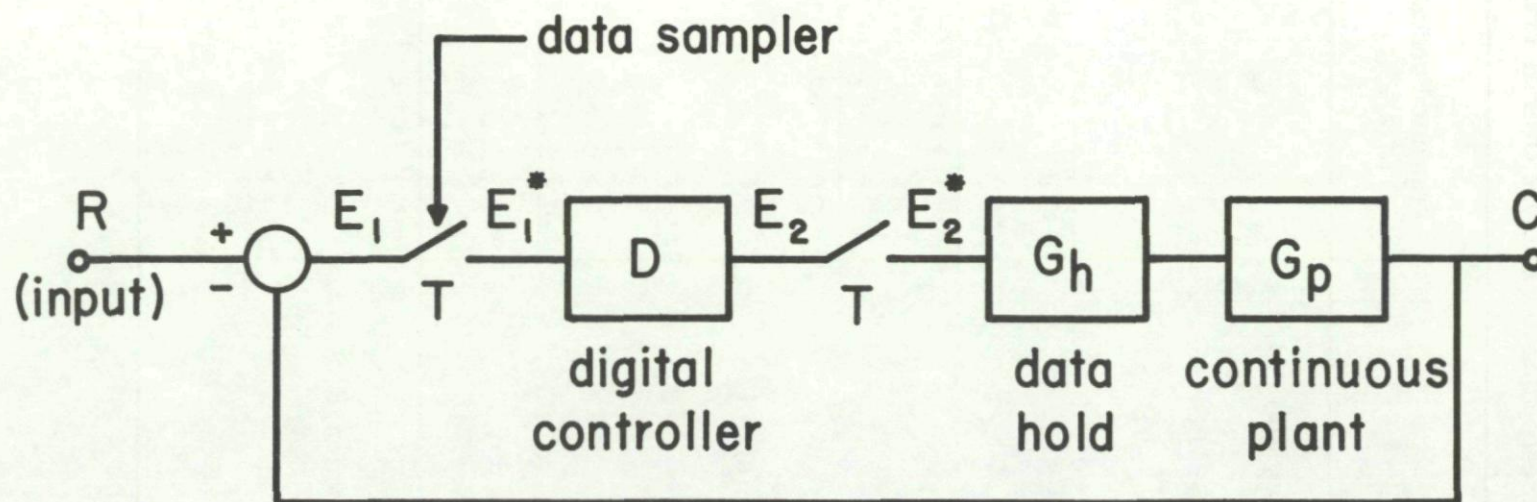


Fig. 1. Typical sampled-data feedback control system configuration.

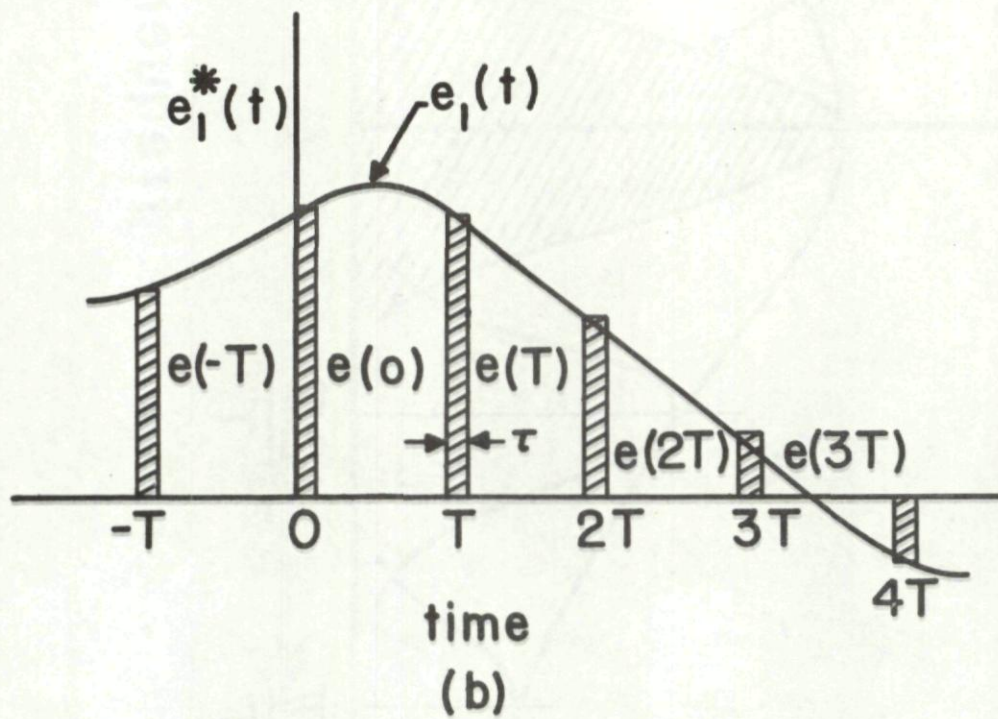
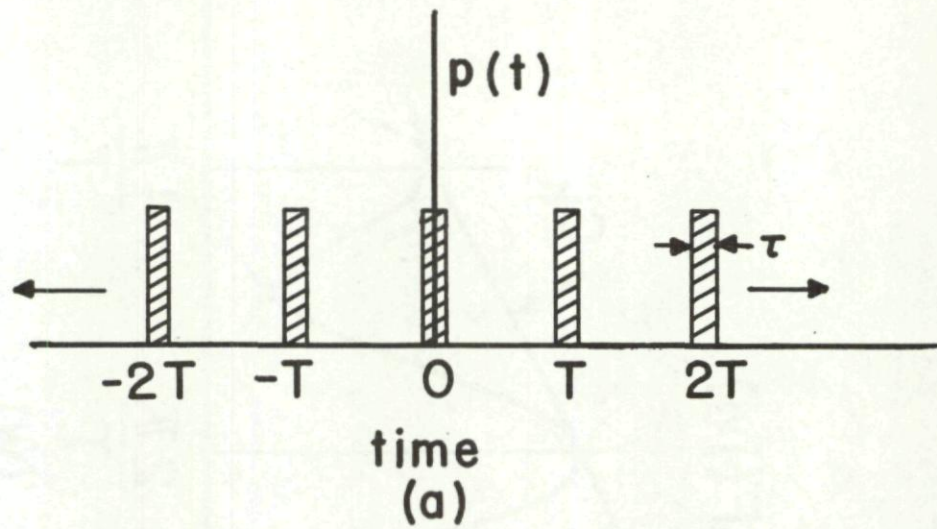


Fig. 2. (a) Periodic sampling function, $p(t)$.
(b) Sampled-signal function, $e^*(t)$.

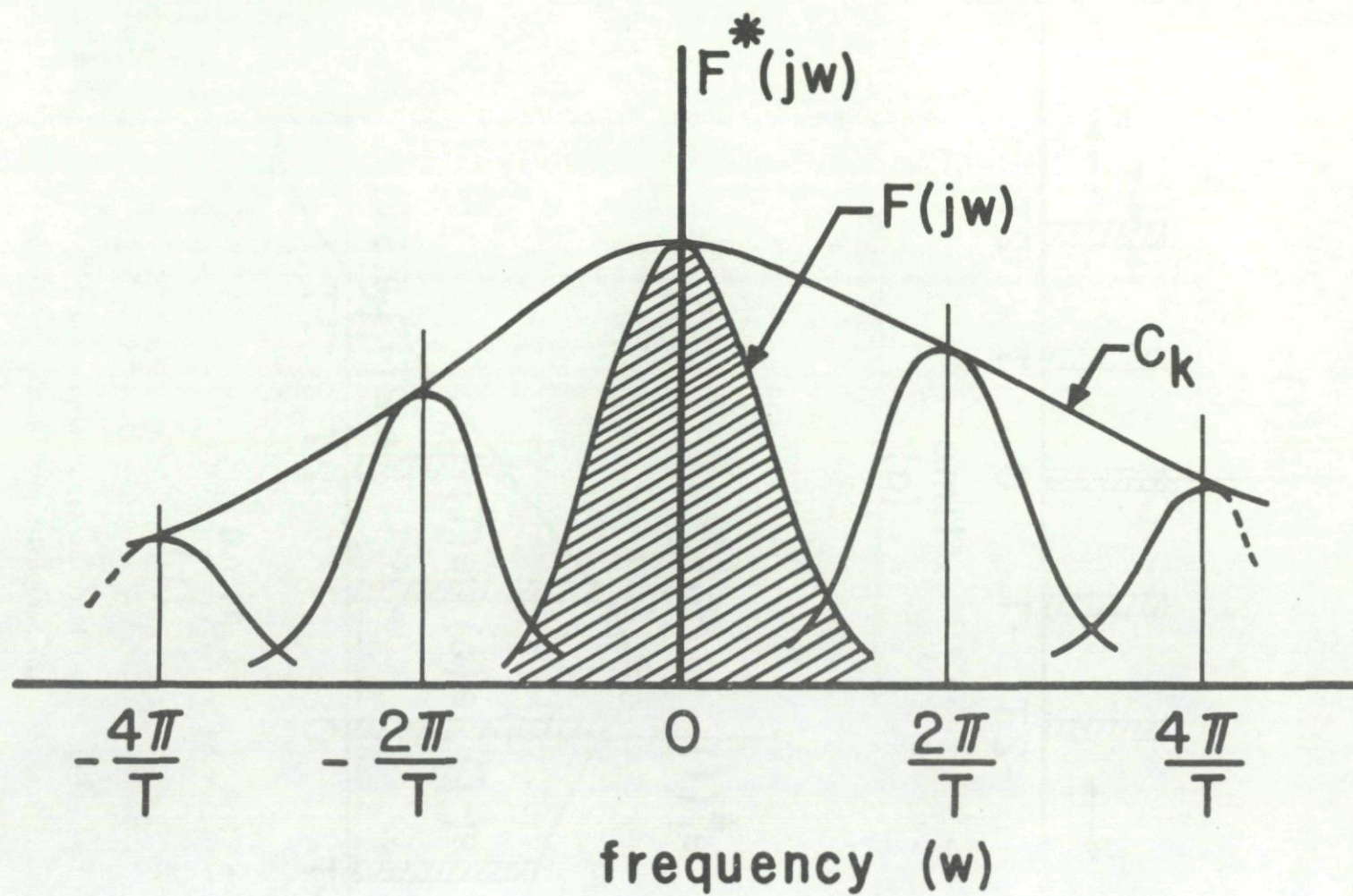


Fig. 3. Frequency spectrum of a sampled signal.

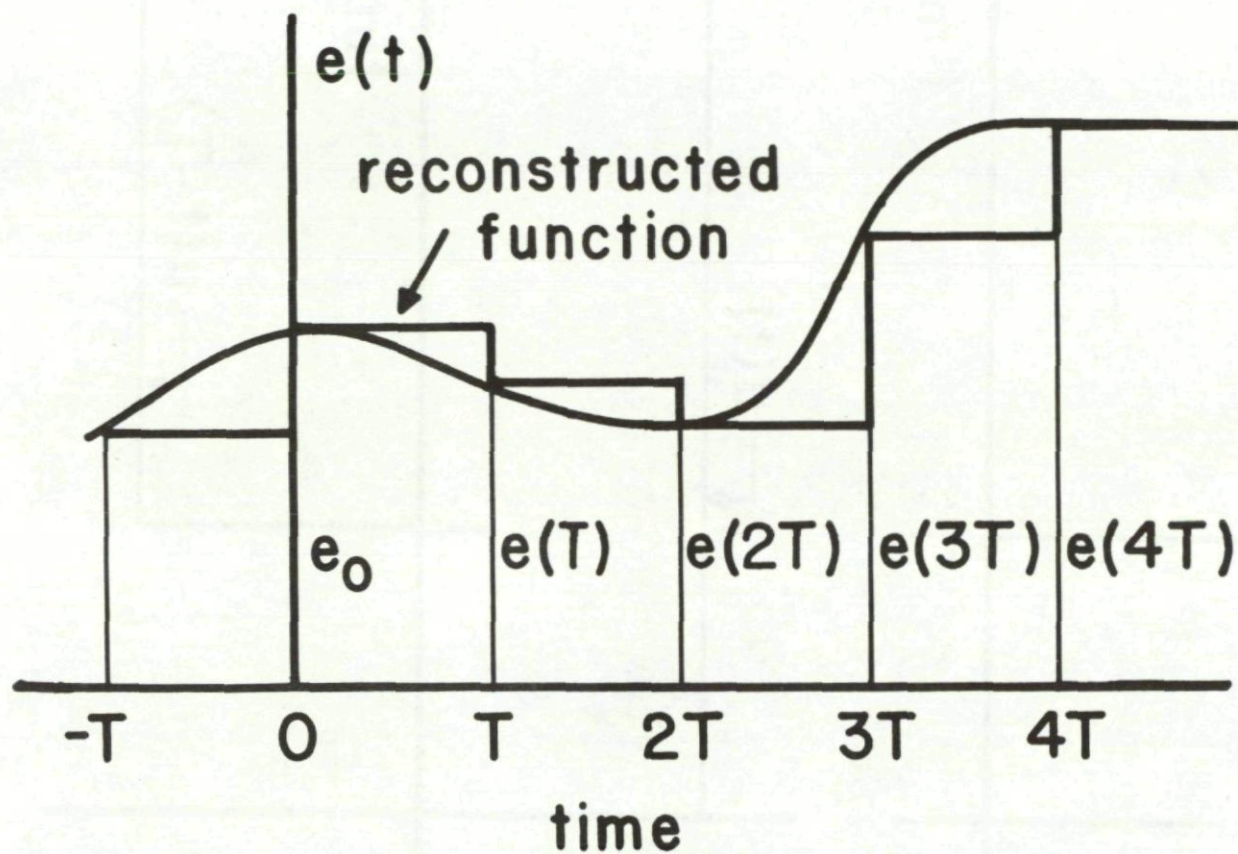


Fig. 4. Operation of a clamp circuit as a data hold.

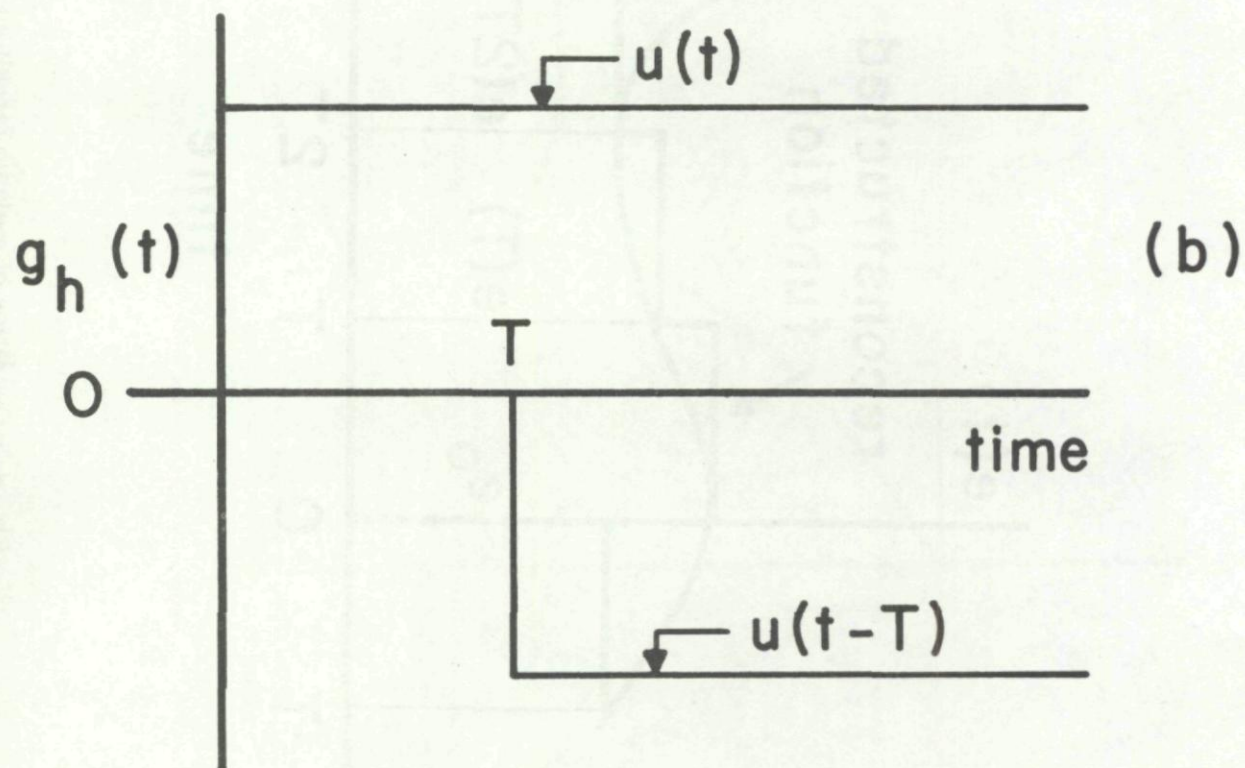
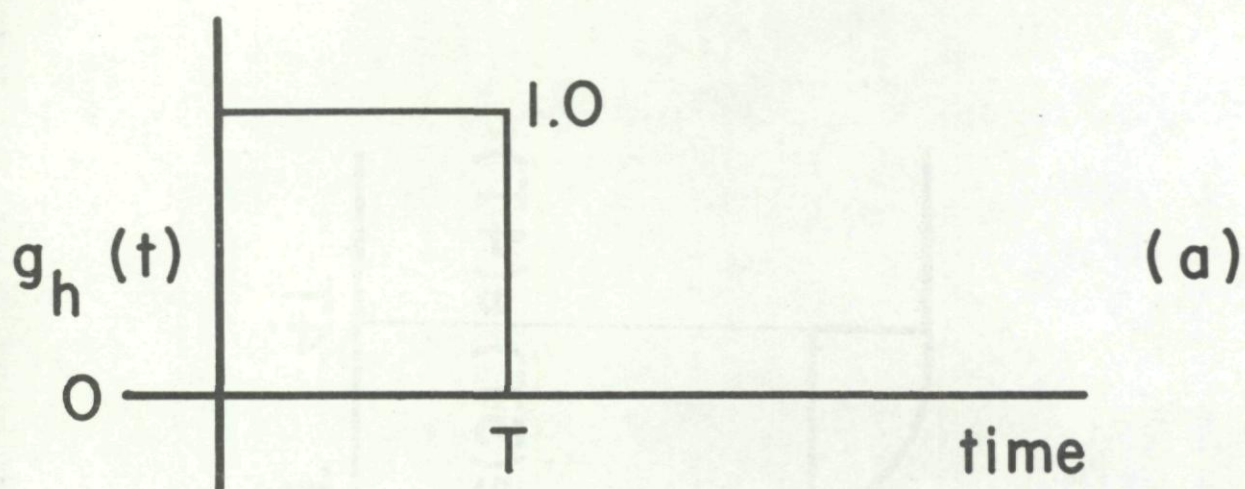


Fig. 5. (a) Impulsive response of a clamp circuit.
(b) Decomposition of impulsive response into two step functions.

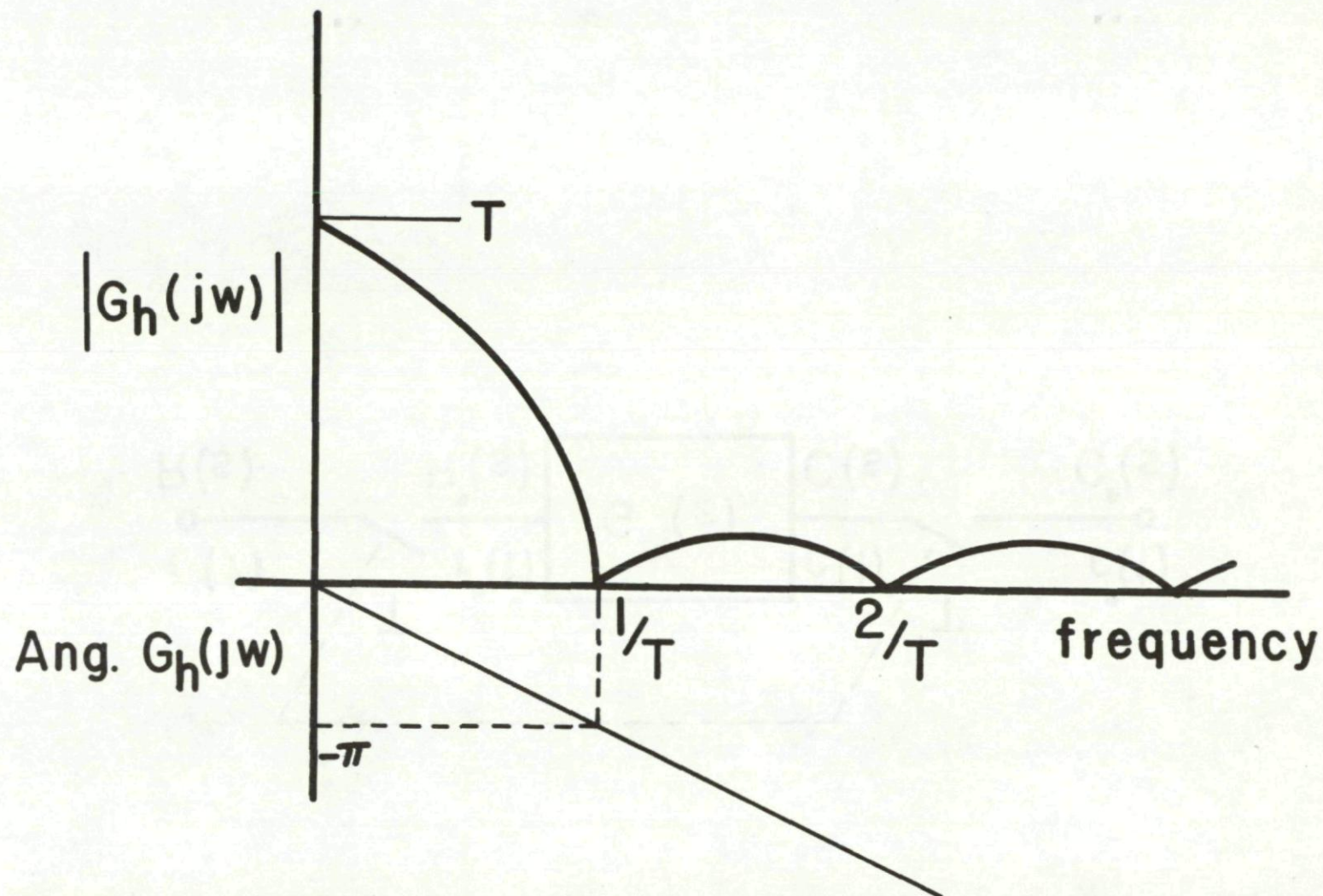


Fig. 6. Frequency response of a clamp (or zero-order data hold) system.

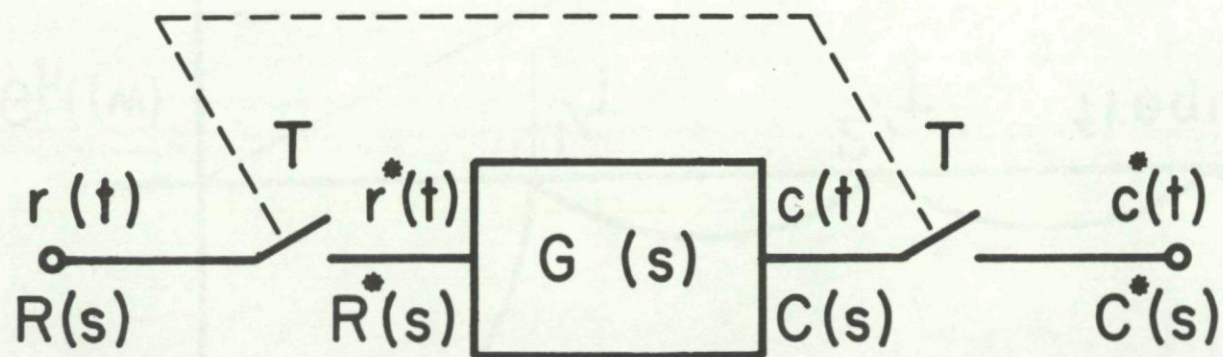
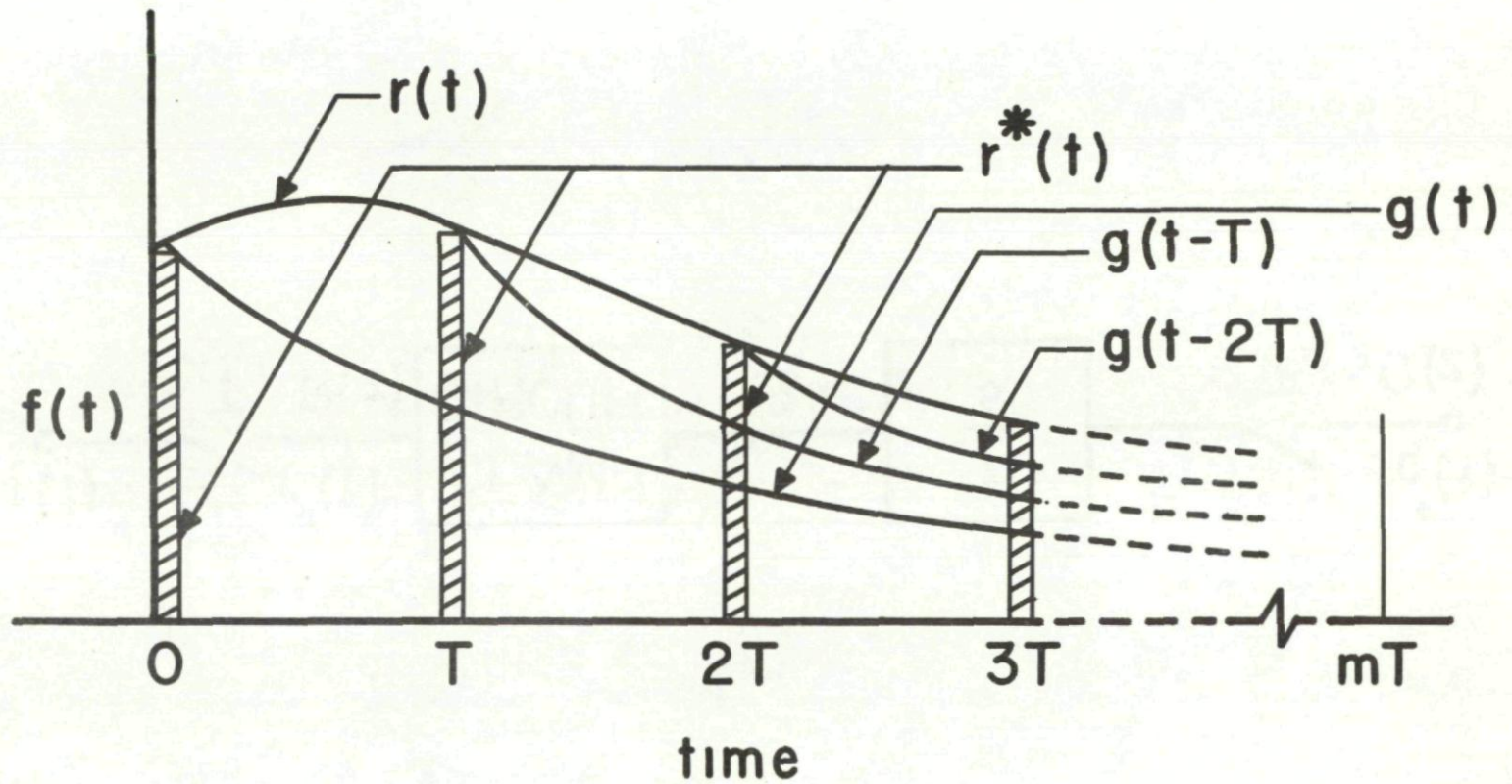


Fig. 7. Typical linear sampled-data system showing inputs and outputs.



8. Combination of impulsive response functions used to obtain the response of system to an input $r(t)$.

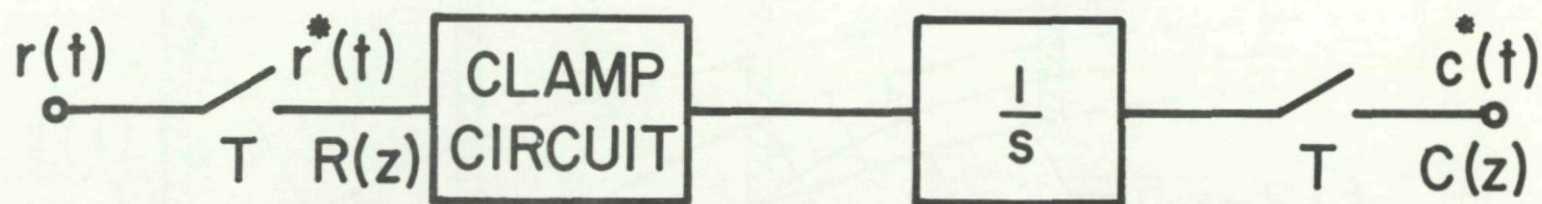


Fig. 9. System used in example to demonstrate use of z-transform method.

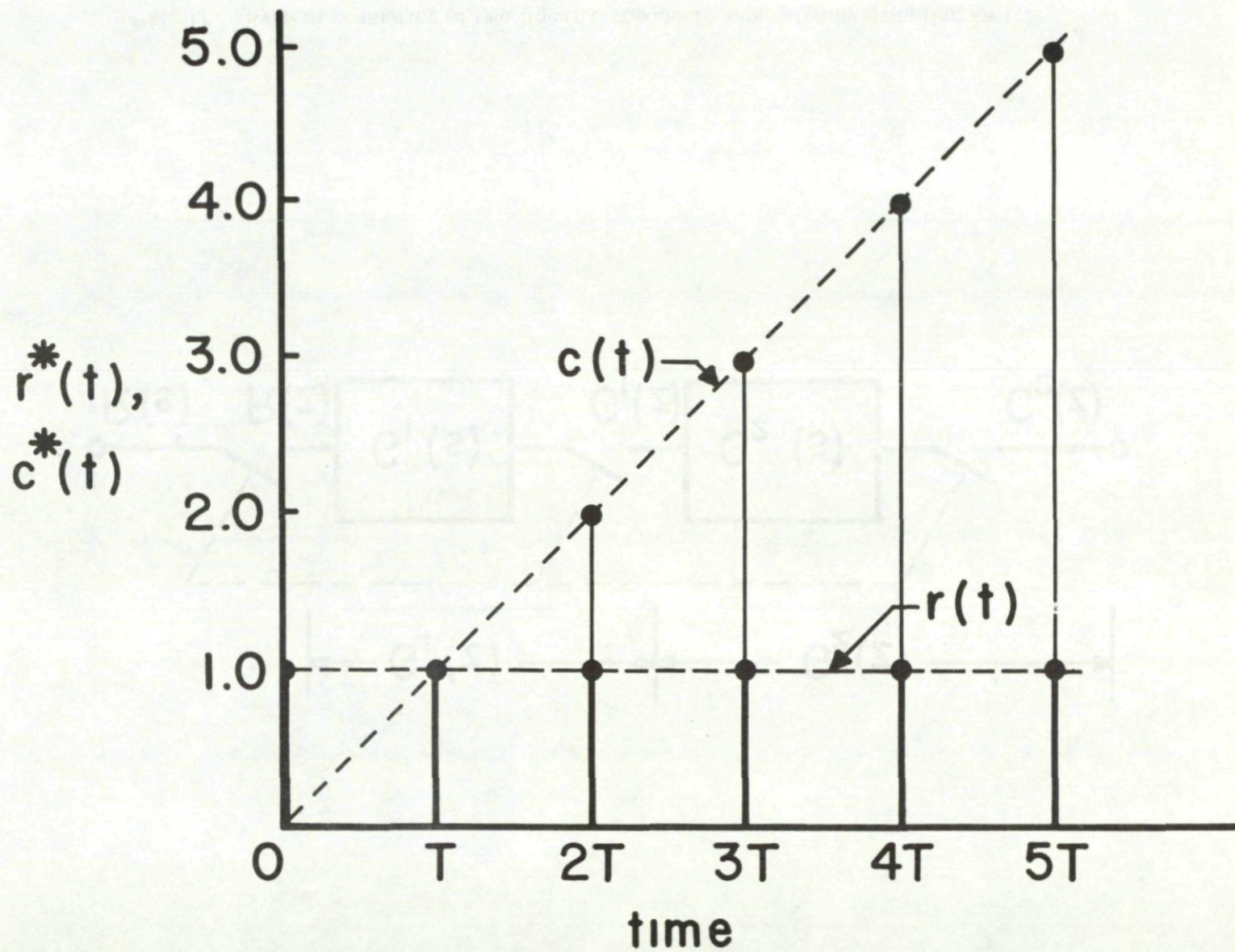


Fig. 10. Output of system shown in Fig. 9 for input step function.

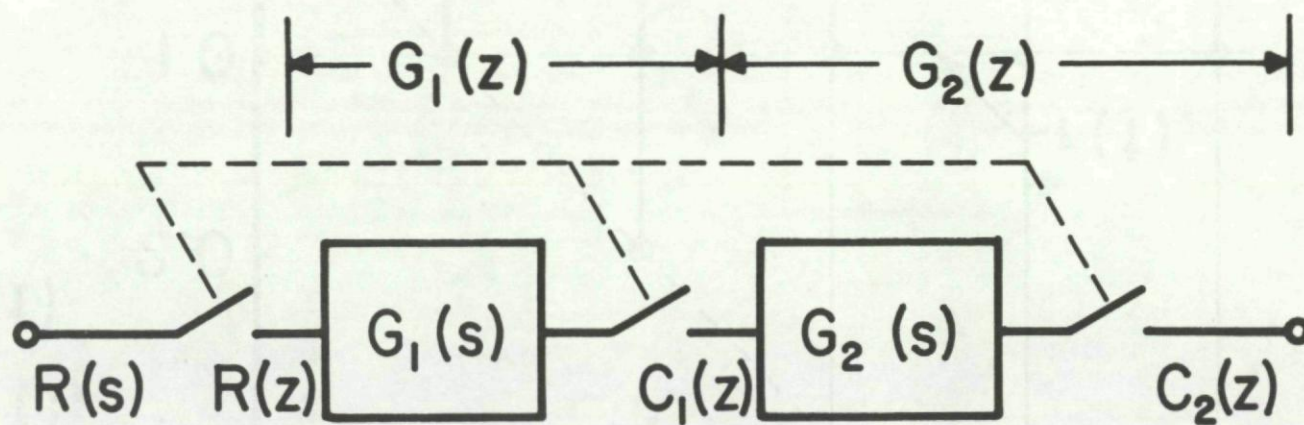


Fig. 11. System consisting of two linear components separated by sampling switch.

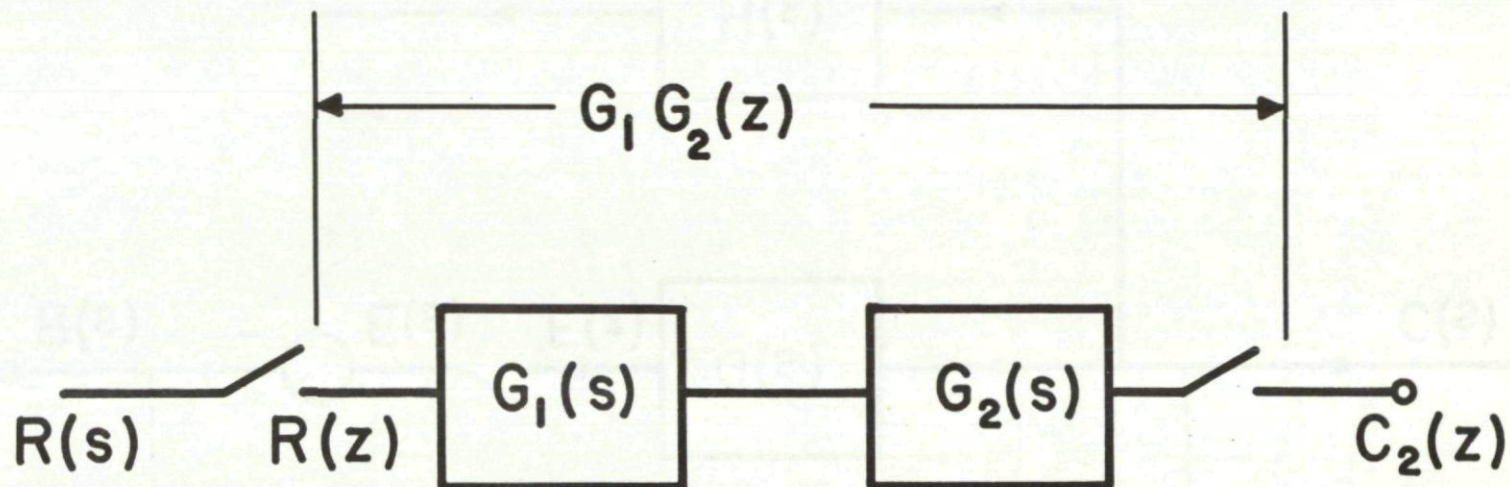


Fig. 12. System consisting of two linear components not separated by sampling switch.

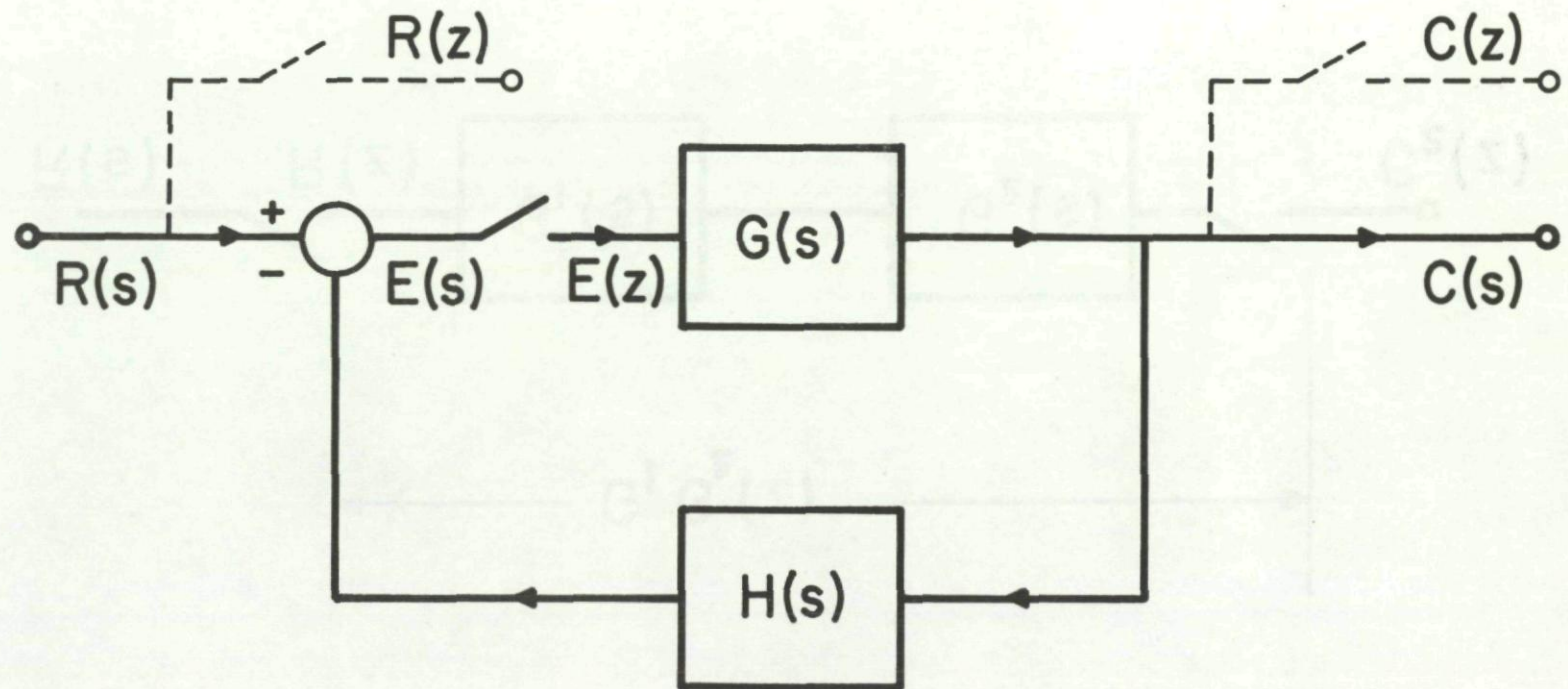


Fig. 13. Error sampled feedback control system configuration.

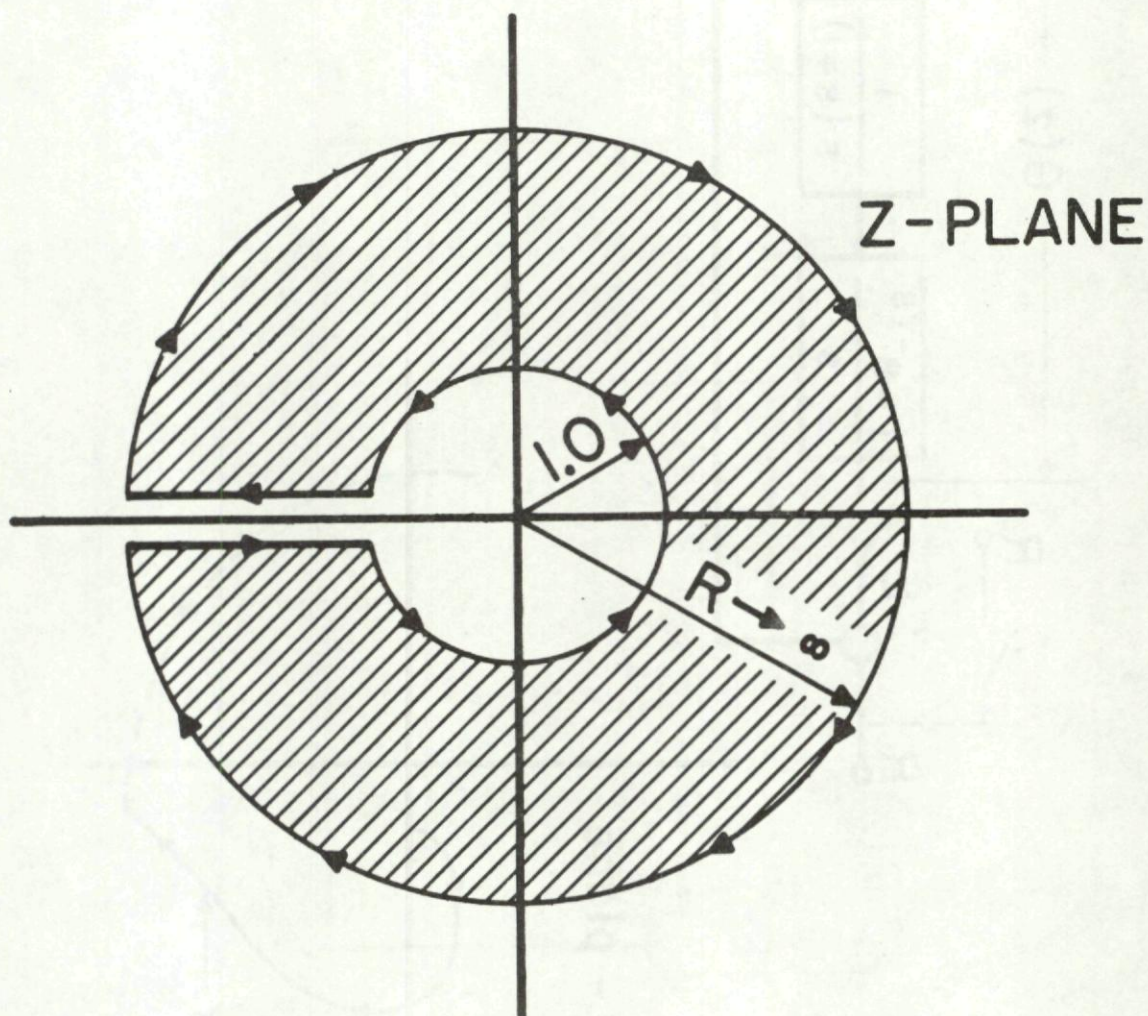


Fig. 14. Locus in z -plane used to map in $G(z)$ -plane for determination of stability.

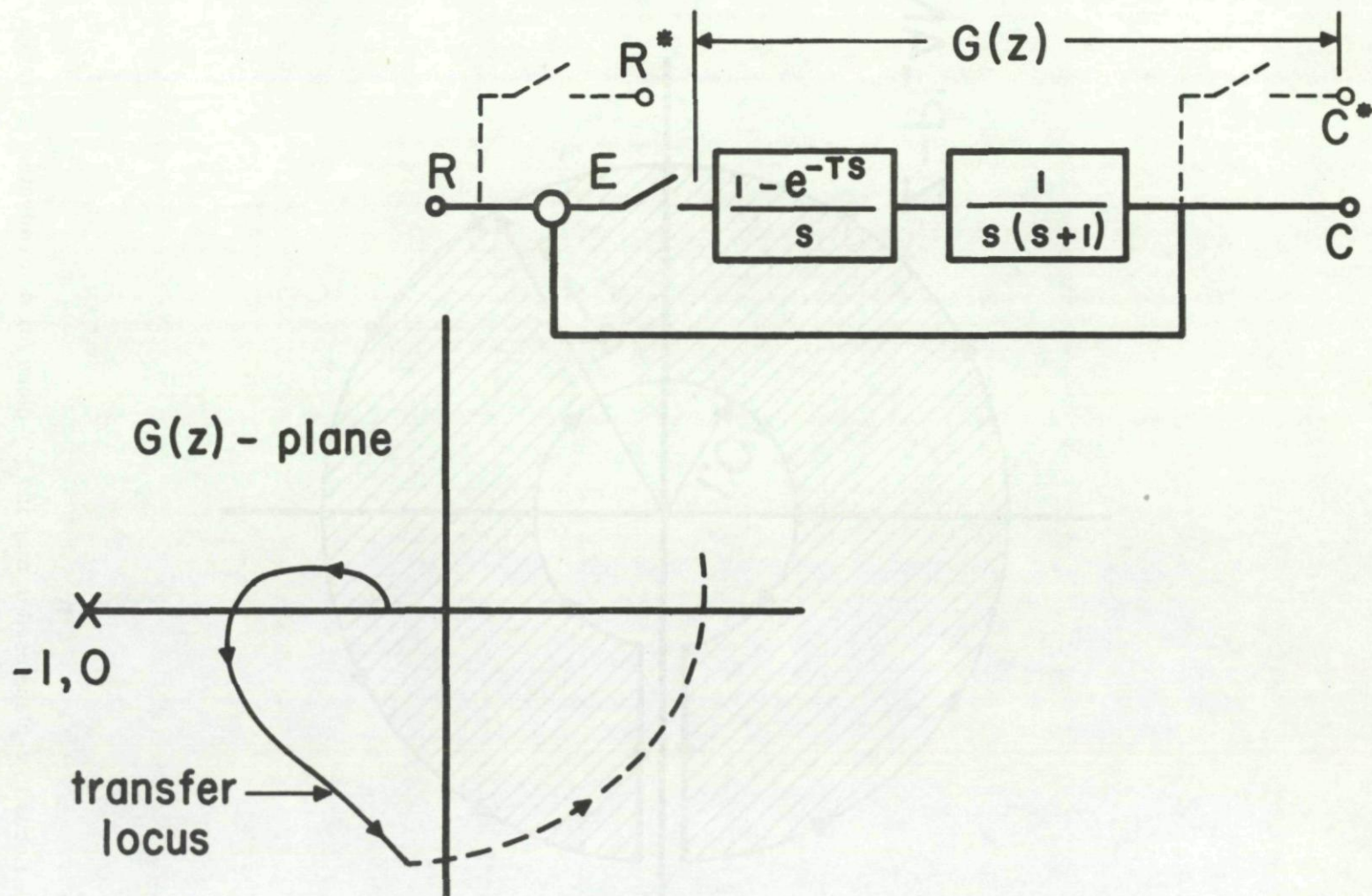


Fig. 15. Transfer locus for typical sampled-data feedback control system.

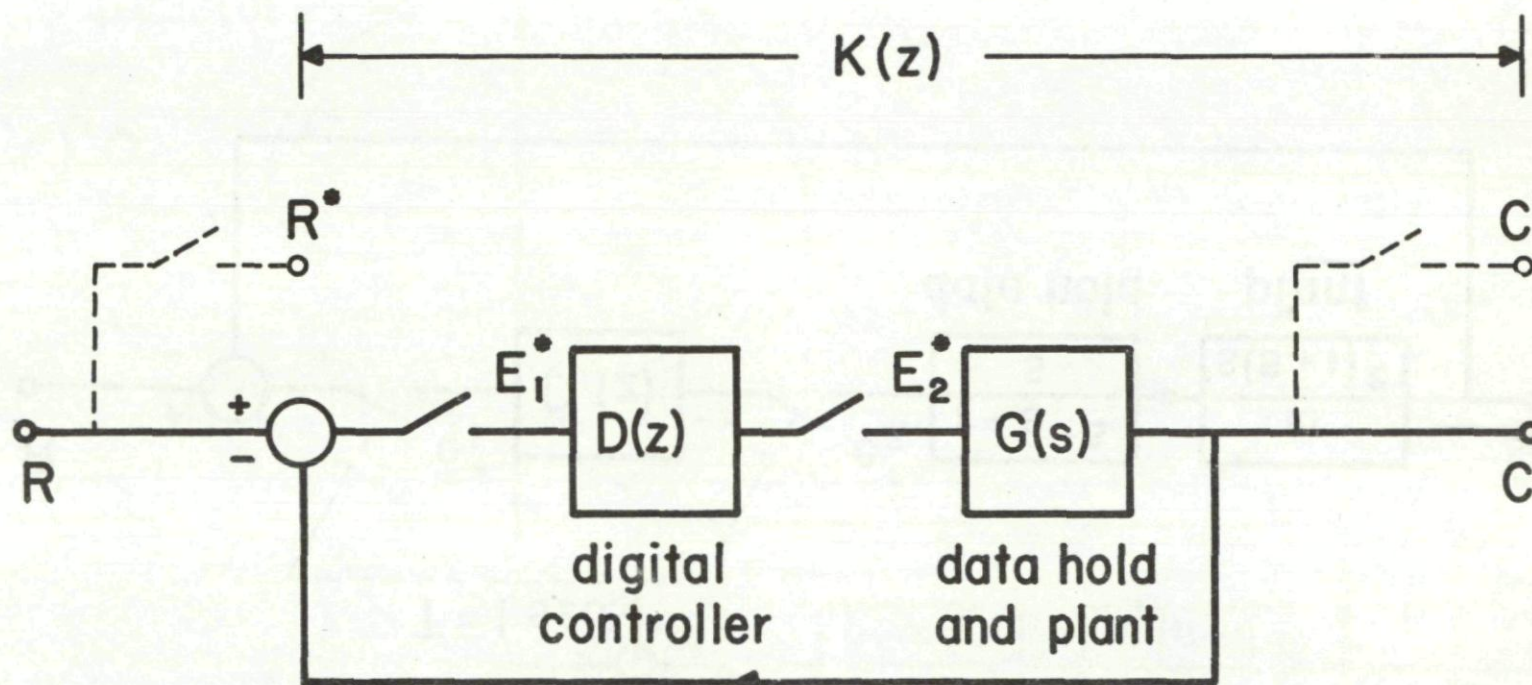


Fig. 16. Typical error-stabilized sampled-data feedback control system.

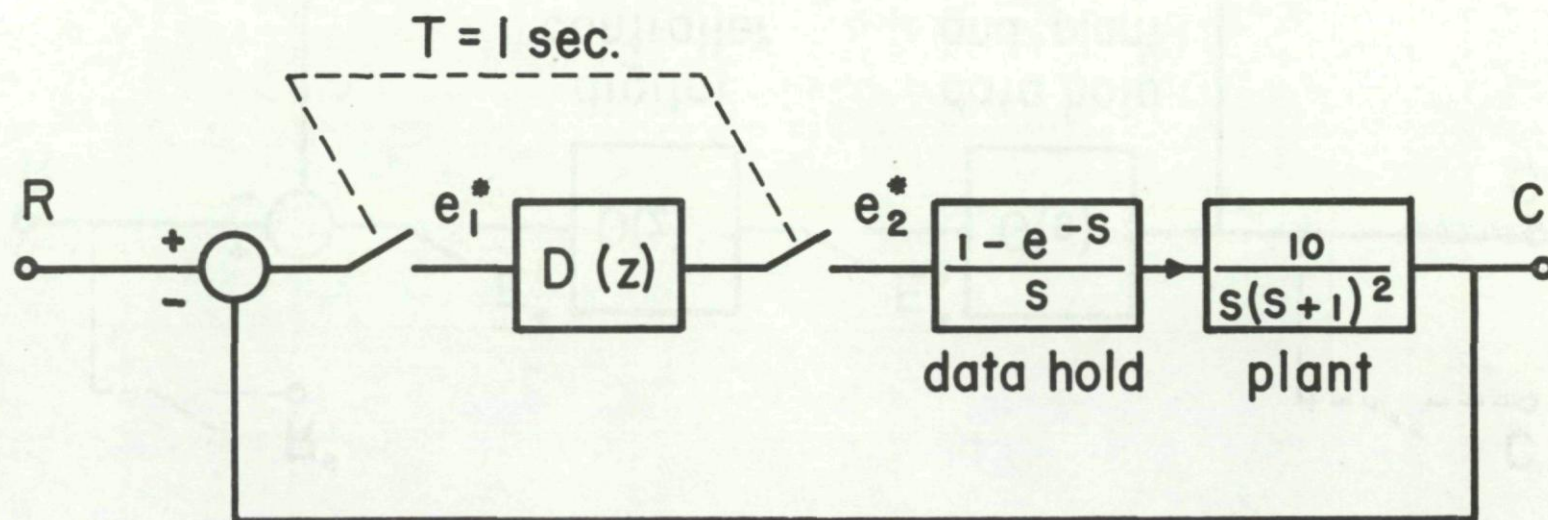


Fig. 17. Typical error-sampled feedback control system used in example.

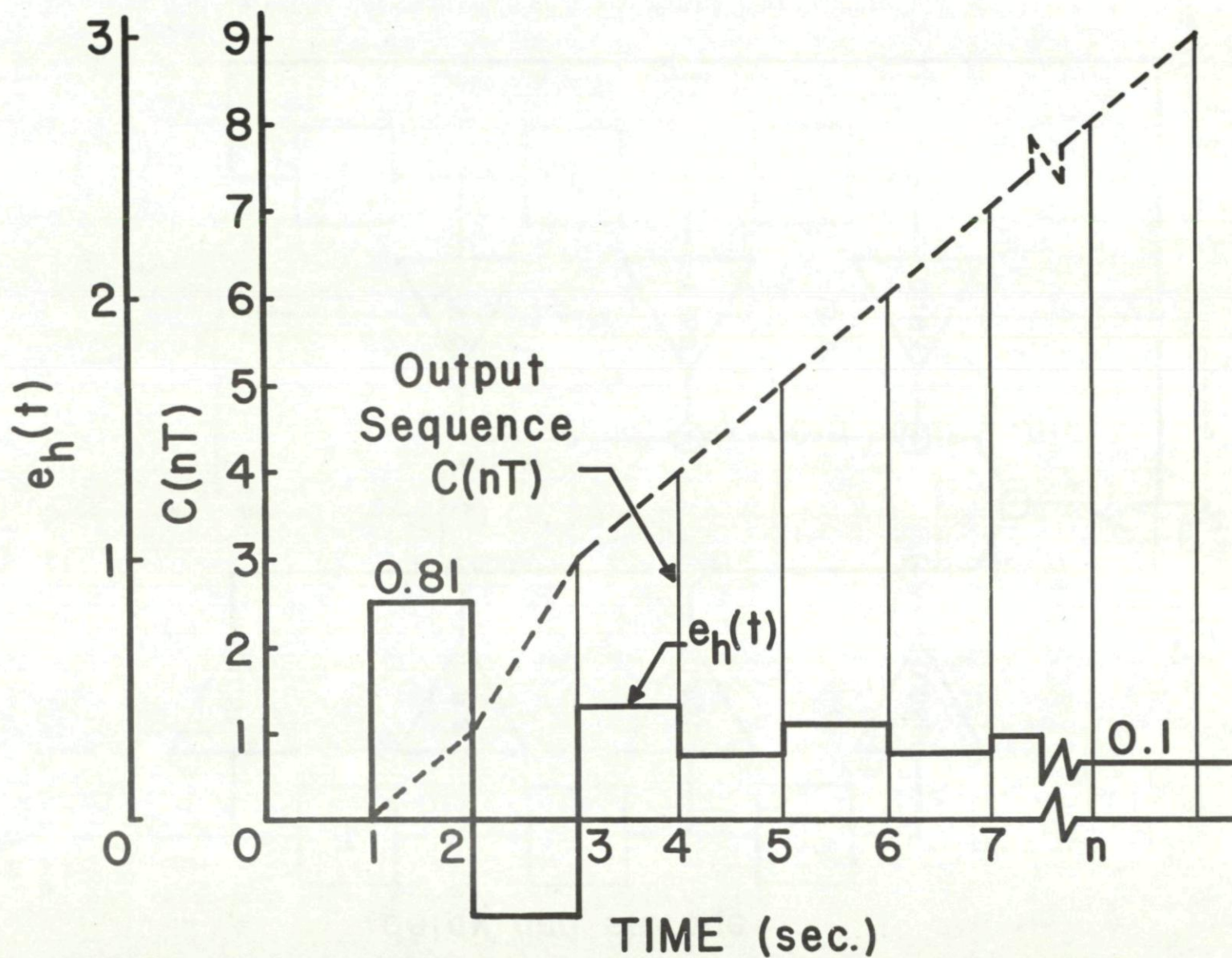


Fig. 18. Output pulse sequence in response to a ramp input for example.

DIGITAL TECHNIQUES IN MISSILE GUIDANCE SYSTEMS

Sidney Darlington*

SUMMARY

This paper describes ways in which digital techniques are useful in guidance and control systems, for missiles and aircraft. It is shown that considerable information or intelligence must usually be incorporated in a guidance system, concerning suitable trajectories, tactics, etc. In addition, numerical information furnished by data collecting instruments must generally be processed in fairly complicated ways. These two functions can be combined in a single digital computer. External and internal characteristics of digital computers, appropriate for guidance applications, are discussed in general terms. Then the programming of typical guidance mathematics is examined, and its relation to the full three-dimensional guidance problem.

SOMMAIRE

Cette note décrit les manières dans lesquelles les techniques digitales sont intéressantes dans les systèmes de gouverne et de contrôle pour les missiles et l'aviation. Il est démontré qu'habituellement une information considérable, ou de l'intelligence, doit être incorporée dans un système de gouverne, en vue d'obtenir des trajectoires, des tactiques, etc, convenables. De plus, l'information numérique fournie par les instruments rassemblant les données, doit être généralement traitée ou interprétée de manière assez complexe. Ces deux fonctions peuvent être combinées dans une seule calculatrice digitale. Les caractéristiques intérieures et extérieures de calculatrices digitales, appropriées aux applications de gouverne, sont traitées en termes généraux. La mise en programme d'équations mathématiques de gouverne typique et sa relation au problème complet de gouverne en trois dimensions sont ensuite examinées.

1. INTRODUCTION

This paper is concerned with the use of digital techniques in automatic guidance and control systems for missiles or aircraft. There are many different digital mechanizations now in existence, of course, with characteristics which differ in many details. There are also many different guidance and control problems, with requirements which differ in many details. For present purposes, however, it will be sufficient to take a very

general point of view. In what ways are digital techniques generally appropriate in guidance and control systems of usual sorts?

Section 2 describes some general properties of guidance and control systems. It shows how guidance systems commonly require information storage (information regarding trajectories and tactics) and number processing (processing of observed missile behavior). The two features can be

*Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

combined in a single digital computer. Other applications of digital techniques are possible, but they are relatively minor and are also relatively specialized.

Section 3 describes external characteristics of digital computers, of sorts appropriate for guidance and control applications. Section 4 describes internal characteristics, in very general terms. Section 5 describes how digital computers can be programmed to perform the mathematics usually required for mechanizing the three-dimensional guidance problem.

2. SOME GENERAL PROPERTIES OF GUIDANCE AND CONTROL SYSTEMS

This section first describes what may be called guidance tactics; it then draws various conclusions regarding characteristics needed in guidance and control systems.

a. Guidance Tactics

Consider first the trajectory of a ground-to-ground missile, directed against a fixed target. With old-fashioned artillery there is no guidance (beyond the gun barrel), yet one achieves accuracies of, say, 1/1000 of the range. This is possible because the unguided trajectories can be predicted to high accuracy. When a ground-to-ground missile is guided, the guidance generally is used to attain accuracies which are not attainable with unguided flight. (The missile's unguided flight may be less predictable than that of an artillery projectile; or the required accuracy may be even greater than that attainable with artillery.) An additional use of the guidance may be to achieve trajectory shapes radically different from the trajectories of free flight as a means of improving propulsion efficiency, reducing aerodynamic drags, etc.

Theoretically, guidance may be applied on either an "open-loop" or a "closed-loop" basis. Generally, however, accuracy considerations dictate closed-loop guidance. The fundamental idea of closed-loop guidance is, of course, as follows: The flight of the missile is observed by suitable data collecting instruments. If it departs significantly from a suitable trajectory leading to the target, guidance orders direct it onto a trajectory which does lead to the target. Since many trajectories lead to the target, this may or may not be a single trajectory, fixed in advance.

The degree of arbitrariness is illustrated in Fig. 1. The missile is started on trajectory 1 from the launcher at M_1 to the target at T. Deviations in the propulsion or control system cause it to wander off trajectory 1 and it is observed by the guidance system at point M_2 .^{*} If allowed to continue without guidance, it will miss the target by a wide margin. The guidance system, however, steers it onto trajectory 2, which passes through the target. It would be possible to design the guidance system to steer the missile back onto trajectory 1. This is not necessary, however, and it may be inefficient.

In the case of a long range ballistic missile, the guidance ends at the end of propulsion. By the end of the propulsion phase, the guidance system must have steered the missile onto a free flight trajectory to the target. With a shorter range, aerodynamically steered missile, guidance can be applied even when the missile is close to the target. This does not mean, however, that flight path deviations can be ignored at earlier times. The missile must be kept

^{*}If the "guidance loop" is tight enough, and the observations good enough, the missile will have no opportunities to wander off trajectory 1; but these conditions are not generally realized in practice.

reasonably close to some sort of suitable trajectory at all times so that errors will not accumulate beyond the capacity of the guidance system to make corrections.

When the target is an aircraft (or another missile), the situation becomes even more complicated. The actual point of impact will now depend, of course, on the motion of the target from present time until impact and the guidance system must make a suitable prediction of this motion. If the target maneuvers in ways not previously predicted, the prediction for the remaining time of flight must be corrected. Thus a maneuvering target calls for a continuous adjustment of the trajectory along which the missile is steered, even though its own free flight is perfectly predictable.

The situation is illustrated in Fig. 2. The missile is at point M_2 when the target is at T_2 . The three missile trajectories are then appropriate, respectively, for the three different target flight paths.

It is a function of the guidance system to figure out the tactical situation and to guide the missile accordingly. Within limits, the prediction of target motion can be fairly crude since continuous corrections are possible almost up to impact. The limits, however, depend on the maneuverability of the missile relative to that of the target. If the margin is small, fairly sophisticated predictions may be in order.*

b. Information Storage and Number Processing

It is clear that the guidance tactics will, in fact, be followed only if suitable information, or "intelligence," is stored within

the guidance system. If trajectories are to be much different from free trajectories, the guidance system must store enough intelligence for the generation of trajectories of the desired sort. Even if free trajectories are suitable, except for perturbations, the guidance system must know, at each instant, whether observed conditions correspond to an acceptable free flight trajectory. In an anti-aircraft system, additional intelligence must be stored to enable the missile to counter target evasion.

In order to apply the built-in intelligence to the observed situation, there must be means for processing the quantitative data collected by observation, or data collecting instruments.

In many guidance systems, the guidance tactics depend on both positions and velocities. Usually, the data collecting instruments give only positions or velocities. Then the guidance system must either differentiate the positions or integrate the velocities. When positions are differentiated, filtering or data smoothing is usually required to reduce effects of fluctuations in observational errors.* Thus, integration, differentiation, filtering, and data smoothing are likely to be important data processing operations in guidance systems.

Transformations of coordinates are another sort of data processing which are also likely to be important. Positions or velocities are usually observed in terms of a coordinate system which is convenient for the data collecting instruments. On the other hand, the steering instructions derived by the guidance system must refer to a coordinate system appropriate for the steering means.

*One starting point for a study of impact point prediction is the prediction theory of Wiener and Kolmogoroff. (See Ref. 1.)

*Data smoothing is a fundamental part of the theory of prediction referred to above.

Vector positions observed by a radar, for instance, are usually measured as combinations of slant range, elevation angle, and azimuth angle. A frame of reference of this sort is illustrated in Fig. 3a. On the other hand, steering is accomplished by setting up accelerations in directions determined by the yaw and pitch steering means. Thus, steering instructions generally are referred to a coordinate system like that shown in Fig. 3b.

Additional coordinate systems may also be needed to facilitate integrations, data smoothing, and the like. For example, the navigation of an aircraft or an airborne missile may call for the computation of latitudes and longitudes by suitable integrations of missile velocities.

Substantial additional data processing is also likely to be needed, but of sorts not so easily codified. For one thing, the tactical intelligence actually is stored as mathematical formulas to be applied to the observed data. Other data processing may have to do with the generation of displays used in monitoring the guidance.

Information storage and number processing can be combined in a digital computer. Before choosing a digital computer for this purpose, however, one must make sure that it is fast enough. Actually, the speed requirements are rather modest because of the way in which various feedback processes interact within the overall guidance and control system.

c. Feedback Paths

The closed-loop guidance system exhibits, of course, the usual characteristics of "feedback." The guidance orders depend on the flight of the missile as observed, and the flight, as observed, is modified by the

guidance orders. Actually, the typical guidance and control system has additional feedback paths within the overall "guidance loop." This is illustrated in Fig. 4.

The innermost loop represents one or more servo devices which set the actual steering controls (aerodynamic surfaces or jet deflection means). So that actuators need not be accurately calibrated, the actual deflections are measured and are compared with the deflections ordered.

The deflections of the steering controls are ordered in accordance with a second feedback path. Here missile orientations or attitude angles are fed back. (Orientations may be measured either with respect to the aerodynamic slip stream or with respect to some fixed reference direction. They may be measured either in degrees or in terms of accelerations produced by the deflections.) The orientations fed back are compared with orientations asked for by the guidance system itself. The differences produce deflections of the steering devices, such as to keep the differences small.

The two inner loops, together, are commonly called the autopilot. The orientations set up by the autopilot, together with the speeds set up by the propulsion, determine how the missile actually flies. In other words, they determine what positions and velocities are actually achieved by the missile. These are observed by the guidance system, and are used in ordering further changes in the orientation.

Very roughly, the various feedback loops are interrelated in the following way: The inner loop is simply an electromechanical servo for setting a shaft position. It has a relatively short time-constant. The orientation of the missile, on the other hand, corresponds to some sort of integration

of the deflections of the steering controls (since turning moments produce only angular accelerations). (The exact form of the relationship depends upon whether the steering is accomplished by aerodynamic surfaces or by jet deflections.) Velocity components and directions of vector velocities correspond to some sort of integration of the missile orientations (since the orientations determine only accelerations). Finally, positions correspond to integrations of velocities.

One result of this situation is as follows: Autopilot errors do not lead to significant trajectory errors unless they are left uncorrected for an appreciable time. As a corollary, guidance need not be continuous nor need it use position and velocity data which are strictly up to date. It is only necessary that autopilot errors will not integrate into significant guidance errors during the effective lag times.

As an example of the effect of temporary acceleration errors, consider the following: An acceleration error of as much as $1g$, lasting for $1/2$ second, produces a velocity error of 16 feet per second. Suppose the velocity error is observed at the end of the $1/2$ second, and is wiped out by applying a net acceleration of $-1/2g$ during the next $1/2$ second. The position error accumulated during the full 1 second of the velocity perturbation is only 8 feet.

From an accuracy standpoint, then, a guidance computer need furnish steering instructions only intermittently at a rate of perhaps once every 0.1 to 0.5 second. We shall see (in Section 5) that this is a reasonable rate of computation for digital computers in most guidance applications. One must also consider the stability of the overall guidance loop, but it turns out that delays of the order of 0.1 to 0.5 second can usually be accommodated.

3. EXTERNAL PROPERTIES OF DIGITAL GUIDANCE COMPUTERS

We have seen that information storage and data processing are important functions performed within most missile guidance systems, and we have noted that these functions can be combined in a digital computer. This section describes some of the external characteristics of digital guidance computers. Section 4 describes some appropriate internal arrangements by means of which the external characteristics can be achieved.

In actual applications, some of the autopilot functions may also be included in the digital operations performed by the digital computer. As a matter of fact, in an integrated system the line between autopilot and outer guidance loop may be somewhat fuzzy. The remarks of this section still apply in a general way, however, even though they are here directed explicitly at the guidance loop itself.

a. Real Time Operation

The guidance computer is of course a "real time" computer. It must keep up with events as they occur. We have already examined computing speed requirements. We will examine computing speed capabilities in Section 5, and will find that they are likely to be adequate if a reasonably high-speed electronic computer is used.

b. Sampling Interval

The digital computer operates on a cyclical basis. It takes in new data cyclically, computes new steering instructions on the basis of the new data, and delivers the new instructions to the autopilot. In the simplest arrangement, everything is repeated exactly once

during each computing cycle. Then the machine takes in new data exactly once each cycle and puts out new instructions exactly once each cycle. Measurements made at intermediate times are ignored.

The effect of the finite computing interval may be compared with "sampling" in analog devices. The effective sampling interval is, of course, the computing interval itself. It is a measure of the effective speed of the machine which we have already discussed and will discuss further in Section 5.

If certain inputs or outputs vary significantly within the overall computing interval a more complicated pattern of cycles may be used to obtain a higher effective speed where speed is needed. For example, several cycles of a simple extrapolation computation may be used to "update" a rapidly varying output quantity several times within each overall computing interval.

c. Dynamic Inputs and Outputs

The function of the guidance loop is to translate observed physical variables into suitable steering instructions. The physical variables are basically analog in nature (positions and velocities). The steering instructions are obeyed in basically analog terms (displacements of steering devices). Thus, some sort of analog-to-digital conversions and digital-to-analog conversions are fundamental to the use of a digital computer.

Most of the sensing or data collecting instruments which are now available furnish analog representations of physical variables. That is, their outputs are usually shaft rotations, analog voltages, or synchro signals representing shaft rotations. These can, in fact, be translated into digital representations, but the translation is likely to be

rather expensive in terms of system complexity. Sensing instruments which measure physical variables directly in digital terms would be a great help, but are not now generally available. The exceptions are generally quite special. (An illustration is the use of an oscillator and cycle counter to measure time intervals.)

Generally speaking, it is possible to convert the analog outputs of the instruments into digital form without any significant loss of accuracy. That is, the digital representation is not significantly poorer than the analog quantity itself. Clearly, however, the accuracy of the analog instrument's output will not be exceeded without recourse to a digital instrument. Similar remarks apply to output conversions from digital-to-analog steering instructions. Roughly, the accuracy is limited by the accuracy to which the analog quantity can be handled in analog operations.

Within the digital computer, of course, computations can be relatively very precise. Accuracies of number processing are limited only by cost and speed considerations.

d. Preset Constants

Within the computer, the trajectory and tactical information is stored in the form of various formulas, to be applied to the dynamic inputs. In many applications constants appearing in these formulas must be adjusted before each firing to coincide with the current tactical situation. In a ground-to-ground guidance system, for example, the trajectory information must take account of the range and direction from the launcher to the particular target which is to be attacked. (In a ground-to-air missile, target location must come in as a dynamic input.) In general, then, it must be possible to supply the computer with a number of constants which are to be different for different shots.

The simplest system may be such that each individual constant is set by hand. If there are more than a few constants, however, reliability considerations are likely to dictate a more automatic system. A punched card system is one possibility, of course, with one card or set of cards recording all constant settings for one target. Cards can be cut and checked for all initial conditions likely to be of interest. Then operator errors can enter only in the selection of cards rather than in the setting of each constant.

e. Displays

In general, certain variables must be displayed before one or more operators so that the performance of the whole missile system can be monitored during the flight of the missile. Various techniques are available for deriving suitable displays from digital quantities within the computer. It can perhaps be said, however, that none of the present techniques are entirely satisfactory. The trouble is that the digital numbers are usually in binary form and must be converted either to decimal numbers or to analog variables before they can make sense to a human operator.

f. Flexibility of Programming

The digital computer does its job by subjecting the input data to a specific sequence of elementary operations. These include the transfer of numbers from one part of the machine to another and the application of ordinary arithmetic operations such as addition, multiplication, etc. It is possible to fix the sequence of operations or program by permanent interconnections. Then the program can be changed only by taking the machine apart and making internal modifications. It is generally preferable,

however, to have a flexible program. It is preferable to have some means of storing any program within some reasonable limits, in a way analogous to the storing of preset constants. One of the important advantages of digital computers relative to analog computers is that they lend themselves more naturally to flexible programming.

When the program is stored in an easily changed way, the guidance program can be replaced by "diagnostic routines" during equipment checks. Changes in the guidance equations reflecting field experience or changes in missile design can be introduced without rebuilding the computer. Within limits, the one computer becomes a "general purpose guidance computer" applicable to different guidance systems without modification. The limits are set by a number of "dimensions" such as capacities for storage of program steps, preset constants, and dynamic variables; number of inputs and number of outputs; number of digits (determining computation accuracies), computing speed, etc.

g. Reliability

A guidance computer must be reliable in a way not usually required of a general purpose laboratory computer. If a laboratory computer makes a mistake the work can usually be done over again. It is only necessary to know when mistakes have, in fact, been made, and to be able to afford the machine time necessary for repeating the computations. If a guidance computer makes a mistake, however, it may misguide the missile in a way which spoils its chances of hitting the target. Then the missile itself is wasted as well as the computing time of the guidance computer.

Accordingly, the design of digital computers for missile guidance is very strongly influenced by reliability considerations.

h. Computer Environments

Computer environments are quite different in different guidance application. In a "command" system in which a missile is guided from the ground, the computer may be in a trailer or even in an air-conditioned building. In an airborne system, it may be inside the cabin of an aircraft. In other systems it may even be inside the missile itself, subject to resulting extremes of acceleration, vibration, and temperature fluctuations. In general, digital mechanizations can be designed for all sorts of different conditions just as analog mechanizations can. If anything, digital techniques should be less sensitive to environmental conditions since precision calibrations are not required.

4. INTERNAL CHARACTERISTICS

From the standpoint of internal operations, there are two quite different forms of digital computers. Both of these are recognized as applicable to guidance problems. One of these may be described as arithmetical, the other as incremental.

An arithmetical machine can form sums, products, etc. of complete numbers in a single computing cycle. An incremental machine, known also as a digital differential analyzer, can only make small changes in previous results. (Even multiplications are performed by sequences of small changes comparable to the mechanical integrations used for multiplication in the earliest differential analyzers.) Incremental machines can be used in guidance systems because the significant numbers do not, in fact, change very much in any one computing cycle.

In general, incremental computers are simpler than arithmetical machines, but they do have some rather serious disadvantages. Also, they are harder to understand. Accordingly, we will restrict our

attention to arithmetical computers. It should be borne in mind, however, that the incremental machines do also exist.

a. Functional Schematic of a Guidance Computer

Fig. 5 illustrates the major parts which are likely to appear in a digital computer as used in a guidance system.

The arithmetic unit does the actual processing of numbers. It adds, subtracts, and multiplies. It may also divide and extract square roots. Generally, it also performs various other operations such as limiting, shifting of decimal (or binary) points, extracting special digits (or binary bits), and testing numbers for signs. All numbers processed may be either positive or negative, and answers are given complete with signs.

In general, all numbers are represented in binary form, except possibly numbers used in outputs to displays. Generally, binary numbers are represented by combinations of electrical pulses, either serially on single wires, or parallelwise on combinations of wires.

The particular operation which the arithmetic unit performs at a given time is determined by a signal from the unit where the program is stored. The same unit controls the sources of the numbers that are operated on, and the disposition of the new numbers generated by the operations. The numbers are moved from one point to another by switching or routing means (computer control circuits) controlled by the program unit. In the constant storage unit and in the dynamic storage unit there are many different "slots" or addresses. A number is taken out of a particular slot, or is stored away in a particular slot, all in accordance with instructions from the program unit.

Numbers in the constant storage are not changed during the guidance of a missile. Numbers can be taken out, but cannot be put in, except in the preparation of the computer before launch. Numbers in the dynamic storage, however are all subject to change. The dynamic storage is simply a "scratch pad memory" where intermediate results are written down.

The separation of the constant storage and the dynamic storage is far from academic. There is a serious question of reliability as well as a difference in accessibility to write-in. When a number in the dynamic storage is recomputed every computing cycle, an accidental change in the number while in the storage unit will probably be corrected during the next cycle. If a constant is changed, however, it will remain in error throughout the rest of that missile flight. As a result, quite different mechanizations may be in order for the constant and dynamic storage units.

As is indicated in the figure, the computer control circuits also receive instructions from the arithmetic unit. First, these may be timing signals telling the program unit when the arithmetic unit is ready for further instructions. In addition, these may include "conditional transfer" instructions. At a particular step in the program, the next step may be either of two alternatives. The choice depends on the result of the previous arithmetical operation as reported by the arithmetic unit.

b. Components

The arithmetic operations may be mechanized as suitable combinations of elementary "logic circuits" of various sorts ("and circuits," "or circuits," "unit delays," etc.). The same is true of the switching or routing circuits. (Transmission routes are determined by logic circuits rather than by switches with mechanical moving parts.)

Elementary logic circuits can be mechanized in various different ways, using either vacuum tubes or semiconductors (diodes and transistors). For guidance applications, however, there are strong arguments in favor of semiconductor circuitry of one sort or another. It is superior to vacuum tube circuitry in at least three ways. It is smaller and more compact, it requires very much less power, and when suitably designed, it is more reliable.

The situation is much less clear in regard to the other major parts of the computer. Magnetic core matrices are attractive for the dynamic storage. The readout of core matrices is destructive, however, and each readout must be accompanied by regeneration if the number read out is to be retained in storage. As a result, core matrix reliability is considered doubtful for the storage of the constants and of the program steps (which must be retained with no errors at all through many computing cycles).

Programs and constants can be stored as interchangeable wire matrices or plug boards, but these become unwieldy in any but quite simple applications. Both programs and constants can also be stored as numbers on magnetic drums or tapes. Actually, drums are in pretty good repute. They do present a timing problem, however, and this may restrict the efficiency of the programming, in a way which is explained in the next section.

Input and output devices for translating between digital and analog variables are many and varied. We will not attempt to describe them in any more detail here.

c. Wound vs. Unwound Programs

Certain sequences of operations are likely to be used several times in a single program. As an example, suppose that the sines

of several different angles are to be computed. An approximate formula will be used, such as a power series, whereby the sine of a given number is constructed out of a series of arithmetical operations applied to the given number and a set of stored constants. Each time a sine is computed, exactly the same process is used except that it is applied to a different number representing a different angle.

A frequently repeated sequence such as that described above is called a subroutine. In some kinds of program storage units a subroutine sequence need not be stored more than once even though it is used many times in the complete program. In other types it must be stored again and again, as many times as it is to be used in the complete program.

In operation a program control unit goes from step to step, or "state" to "state," as the program progresses. At each step or state it issues one set of instructions to the arithmetic unit and the switching circuits. It continues in this manner until the program is completed, and then starts over again in a new computing cycle. If the programmer can advance only one state at a time and can never return to previous states (except when it returns to state one at the beginning of a new computing cycle), a subroutine must be stored once for each time it is used in the complete program. If, however, the programmer can return repeatedly to earlier steps, a subroutine sequence need be recorded in the programmer only for its initial use. For additional uses, the programmer can return to the same sequence of states. In the usual computer terminology, the first programmer can store only "unwound" programs, while the second can accept "wound" programs.

A trigonometric subroutine, which will be described in Section 5, often can be used to take care of a large part of a guidance computer program. Under these circumstances,

it takes much less storage capacity to store the program in wound form than in unwound form. If the program is stored on a magnetic drum, synchronization problems may dictate storage in the unwound form. The storage capacity of magnetic drums is so great, however, that it may be cheaper to store the unwound program on a drum than the wound program in a more versatile mechanization.

5. PROGRAMMING REPRESENTATIVE GUIDANCE MATHEMATICS

We noted in Section 2 that a guidance computer's capabilities must generally include the following: storage of trajectory and tactical information, transformations of coordinates, and computations such as integration, differentiation, filtering and data smoothing. Let us reexamine these in the light of what has just been said about digital computers.

a. Storage of Trajectory and Tactical Information

The information, or intelligence, stored within the computer can usually be represented by functions of one set of variables or another. In practice, the functions are approximated by combinations of arithmetic operations (power series, interpolation routines, and the like). Various techniques are available for programming mathematics of this sort, as a result of the widespread use of general purpose computing machines.

In guidance applications, rather crude approximations are generally permissible when the missile's flight can be refined as it approaches the target. As a result, the corresponding demands upon the computer are likely to be quite reasonable in regard to computing time, program steps, and numbers of constants.

b. Transformations of Coordinates

The two subroutines described below can be used for most or all of the coordinate transformations needed in a guidance computer. Whether they are actually so used in practice may depend upon whether wound or unwound programs are to be stored. In any event, the subroutines give a rough measure of the difficulty of the coordinate transformations in terms of computing time and number of program steps.

The first and more basic subroutine carries out the two-dimensional transformation illustrated in Fig. 6. It is equivalent to a "resolver" in an analog machine. In the figure, x and y are components of a two-dimensional vector R in a given frame of reference. Then x' and y' are components of the same R in a new frame of reference obtained by rotating the x, y frame through a given angle θ . The subroutine determines x' and y' when given x, y , and θ .

Note that θ, R are the components of R in a rectangular frame of reference such that the y axis is along the vector. Thus the same subroutine can be used to resolve a two-dimensional vector into components x and y , when its magnitude and direction angle are given.

The second subroutine performs an inverse sort of operation. It starts with the components x and y and determines the corresponding two-dimensional vector magnitude and direction. One way in which it can be mechanized uses the first subroutine in a feedback or servo loop within the computing program. An angle of rotation ϕ is adjusted in such a way that the first subroutine, applied to x, y , and ϕ , makes $x' = 0$ (and y' positive). Then y' and ϕ are the desired magnitude and direction.

In a simple experimental computer the author used programs of the following lengths, for the two subroutines: for the first subroutine, 53 single address steps; for the second subroutine, a total of 72 (including the use of the first subroutine in a feedback loop).^{*} (This included computation of the necessary trigonometric functions.) If the arithmetic unit had been somewhat more versatile in regard to conditional transfers and limiting, the numbers could have been a little smaller.

For the full three-dimensional guidance problem, sequences of two-dimensional transformations can be used. The following example will illustrate the method: Suppose a target position relative to an aircraft or missile is given as North and East components and difference in altitude, and suppose that the deflection angle, elevation angle, and slant range are desired relative to axes fixed in the air frame. It is assumed that the aircraft heading, pitch, and roll angles are general, but are known. The computation can be accomplished by three operations of the first subroutine, and two of the second.

The above mechanization assumes that heading, pitch, and roll angles are defined as the various gimbal angles of a gimbal mounted stable platform. Air frame angles are likely to be measured in terms of gimbal angles, and gimbal angles are more appropriate for the use of our subroutines than, say, direction cosines.

Another illustration will show the versatility of the two subroutines. Suppose an aircraft or missile is known to be at longitude ψ_1 and latitude θ_1 and that it is to

^{*}In general, it takes two single address steps to add, subtract, multiply, or divide two numbers.

be flown along the great circle course to a target at longitude ψ_2 and latitude θ_2 . What is the bearing of the great circle course at ψ_1, θ_1 , and what is the great circle distance to ψ_2, θ_2 ? This problem also can be solved by three applications of the first subroutine and two of the second. The sequence of operations is derived by representing the destination as components in a rectangular (x, y, z) system of coordinates with its origin at the center of the earth. The problem is solved by suitable rotations of this coordinate system.

c. Integration, Data Smoothing, Etc.

In ordinary guidance computations, integrations are to be carried out with respect to time. They can then be approximated with sums of suitable increments added once each computing cycle.

The integration process is extremely simple except for the possibility of serious accuracy troubles. The accuracy problem arises because small round-off errors in individual increments may accumulate as large errors in the sums. The accuracy problem is an old one, however, and can be solved in various well-known ways.

Differentiation is generally combined with data smoothing. Filtering and data smoothing are usually linear operations representable by linear differential equations.

If the solutions of the differential equations correspond to sampled forcing functions and are themselves sampled at regular intervals corresponding to the computing cycles of the guidance computer, the samples are matched exactly by solutions of difference equations derivable from the differential equations.

The difference equations can be solved by means of a series of integrations corresponding to their various terms. Alternatively, the solutions can be expressed as sums of solutions of systems of first and second order equations of the form illustrated below.

(1) Typical differential equation:

$$\left[\sum_{\sigma=0}^{\mu} a_{\sigma} \frac{d^{\sigma}}{dt^{\sigma}} \right] y = \left[\sum_{\sigma=0}^{\mu} b_{\sigma} \frac{d^{\sigma}}{dt^{\sigma}} \right] x$$

(2) Corresponding system of difference equations:

$$y = y_0 + y_1 + y_2 + \dots + y_s + \dots$$

where

$$y_{0n} = A_0 x_n$$

and either

$$y_{Sn} = A_S x_n + B_S y_{S(n-1)}$$

or

$$y_{Sn} = A_S x_n + B_S y_{S(n-1)} + [C_S x_{(n-1)} + D_S y_{S(n-2)}]$$

The individual first and second order equations can be solved in a routine way. The solution of an n'th order differential equation with constant coefficients required a total of 2n multiplications and from n to 2n additions per computing cycle. It requires n slots in the dynamic memory, and up to 2 n+1 constants.

Table 1. Computing Speed of the TRADIC Phase One Computer

Type of Operation	No. Operations in 1/2 Second
Add or subtract two numbers (including readout from accumulator)	8000
Multiply or divide two numbers (including readout from accumulator)	1600
Rotation (resolver) subroutine (x' , y' from x , y , θ)	100
Magnitude and direction from x , y (one increment each computing cycle)	80

d. Permissible Program Lengths

The complete program may include a substantial amount of additional computing besides the computations described in subsections b and c above. This is too varied in character, however, to be considered further here.

In conclusion, Table 1 is used to show that, in fact, a digital computer can do a lot of computing, within a computing interval

of tolerable length. Table 1 illustrates the number of operations of various sorts which may be performed in a 1/2-second interval. Operations of the subroutines described in subsection a are included, as well as the ordinary arithmetic operations. In practice, of course, the numbers must vary with the specific capabilities of the specific computers employed. The numbers in the figure, however, do correspond to an actual computer, the so-called Phase One TRADIC Computer (Ref. 2).

REFERENCES

1. Bode, H. W., and Shannon, C. E., "A Simplified Derivation of Linear Least Squares Smoothing Theory," Proceedings of the Institute of Radio Engineers, April 1950.
2. Bell Telephone Laboratories, "TRADIC Phase One Summary Engineering Report Volume I," Air Research and Development Command, U. S. Air Force Report No-21536-5, July 1954.

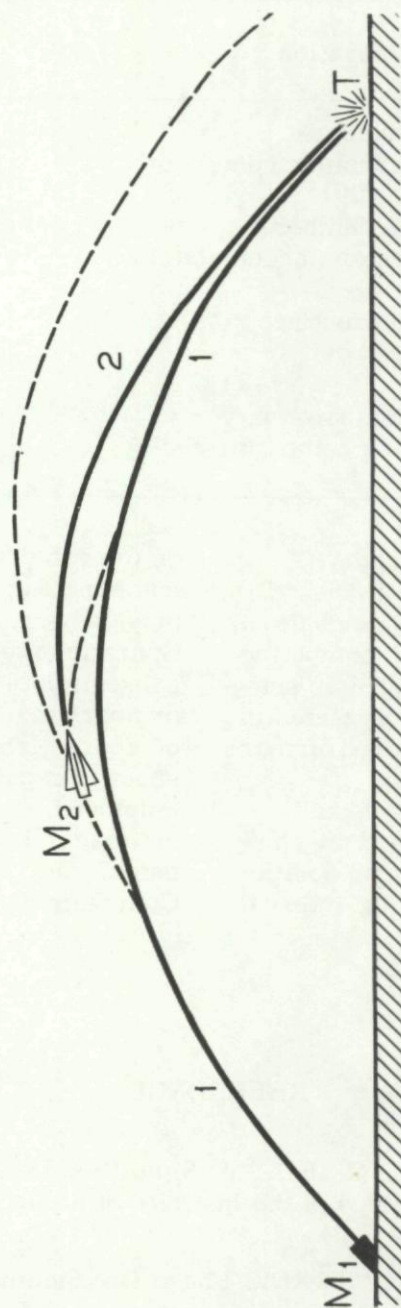


Fig. 1. Ground-to-ground trajectories.

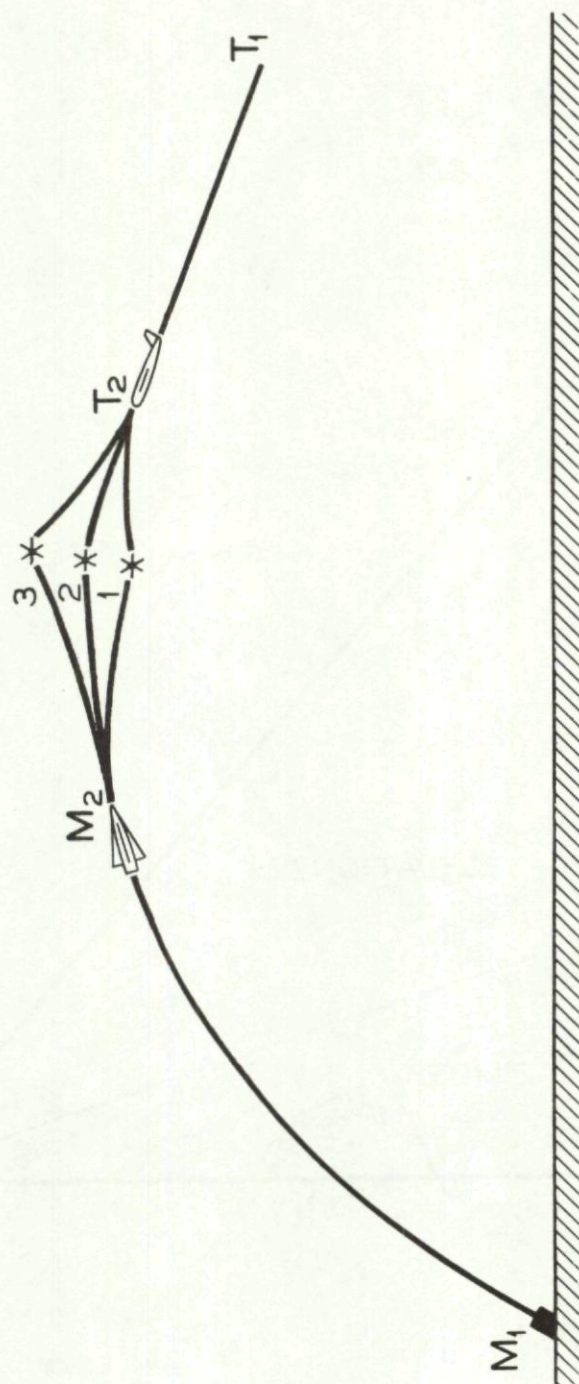


Fig. 2. Antiaircraft trajectories.

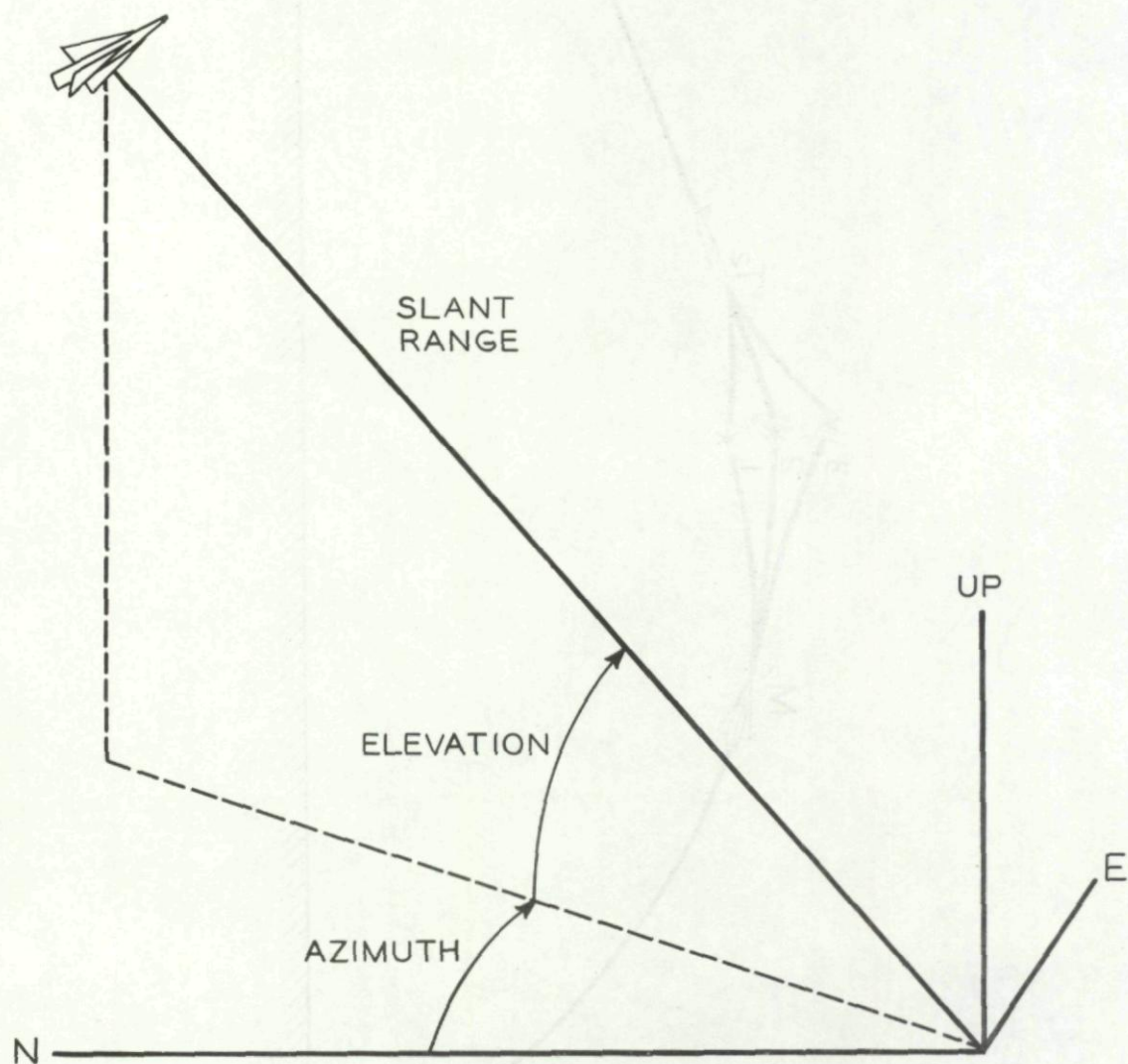


Fig. 3a. Radar coordinates.

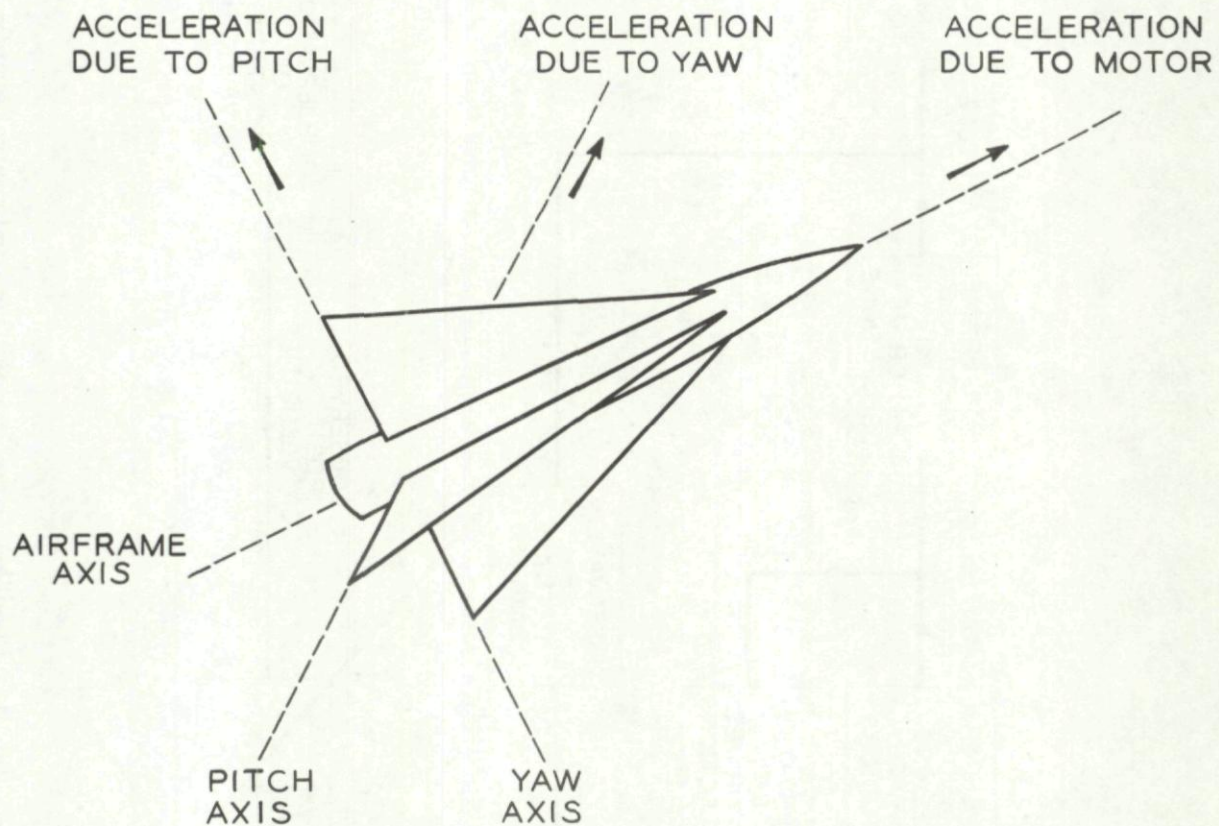


Fig. 3b. Steering coordinates.

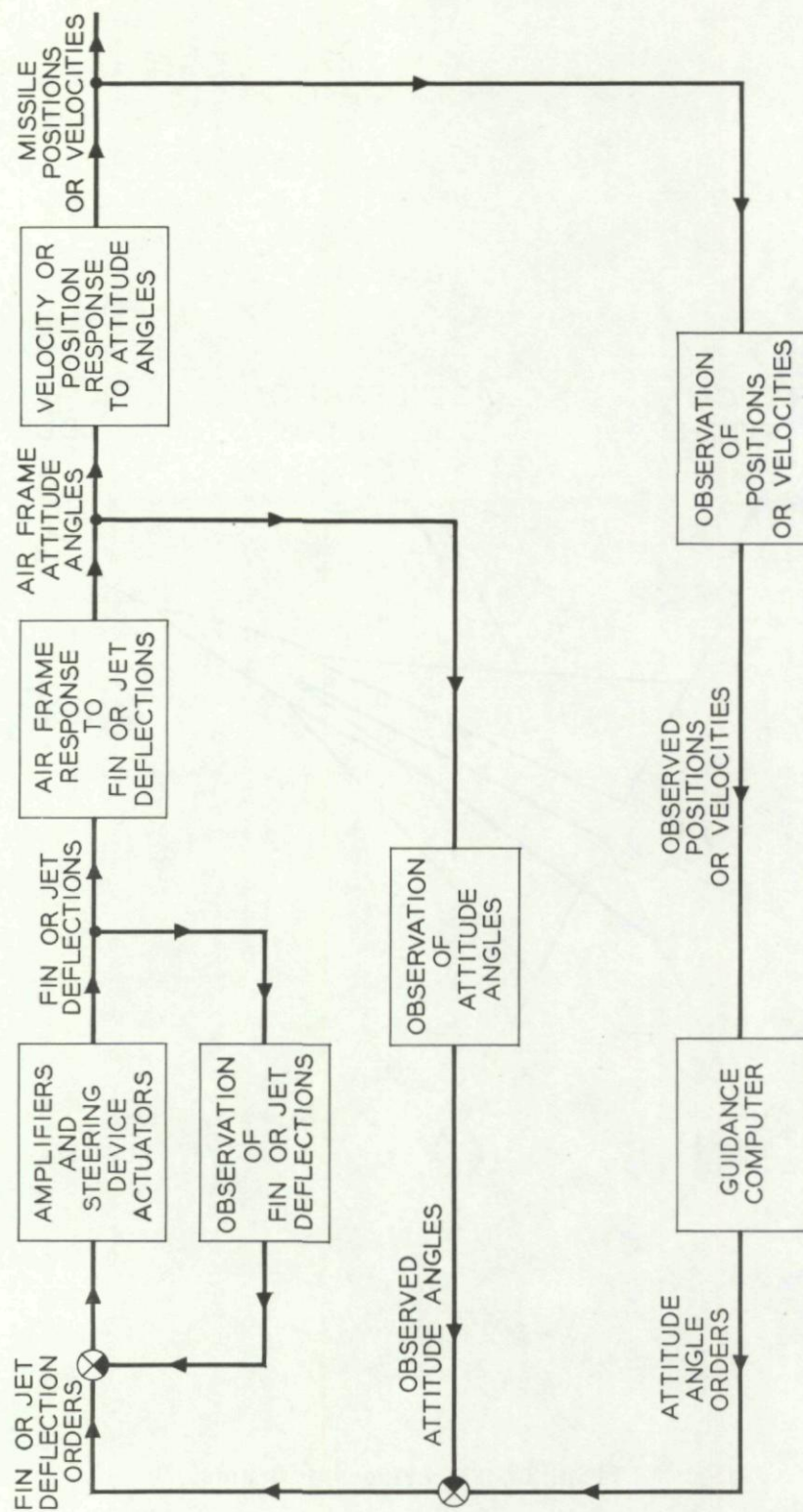


Fig. 4. Feedback paths in a guidance and control system.

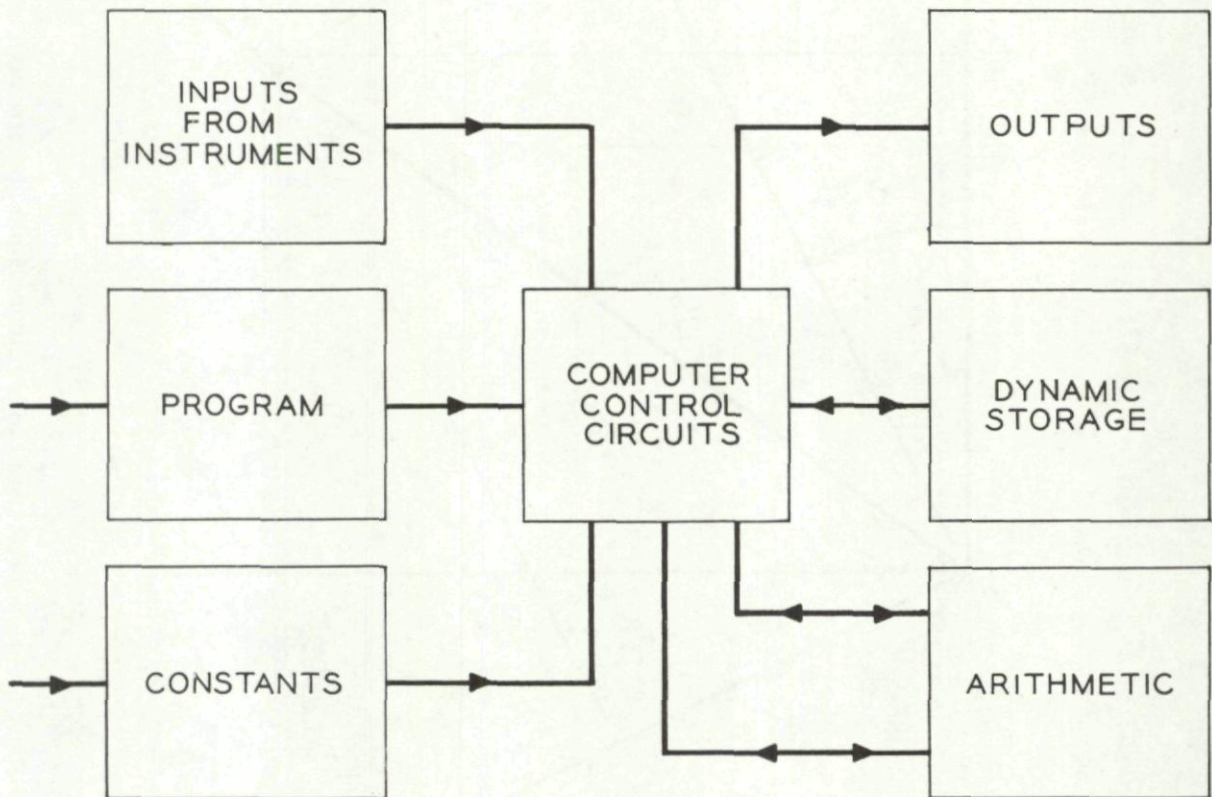


Fig. 5. Block diagram of a guidance computer.

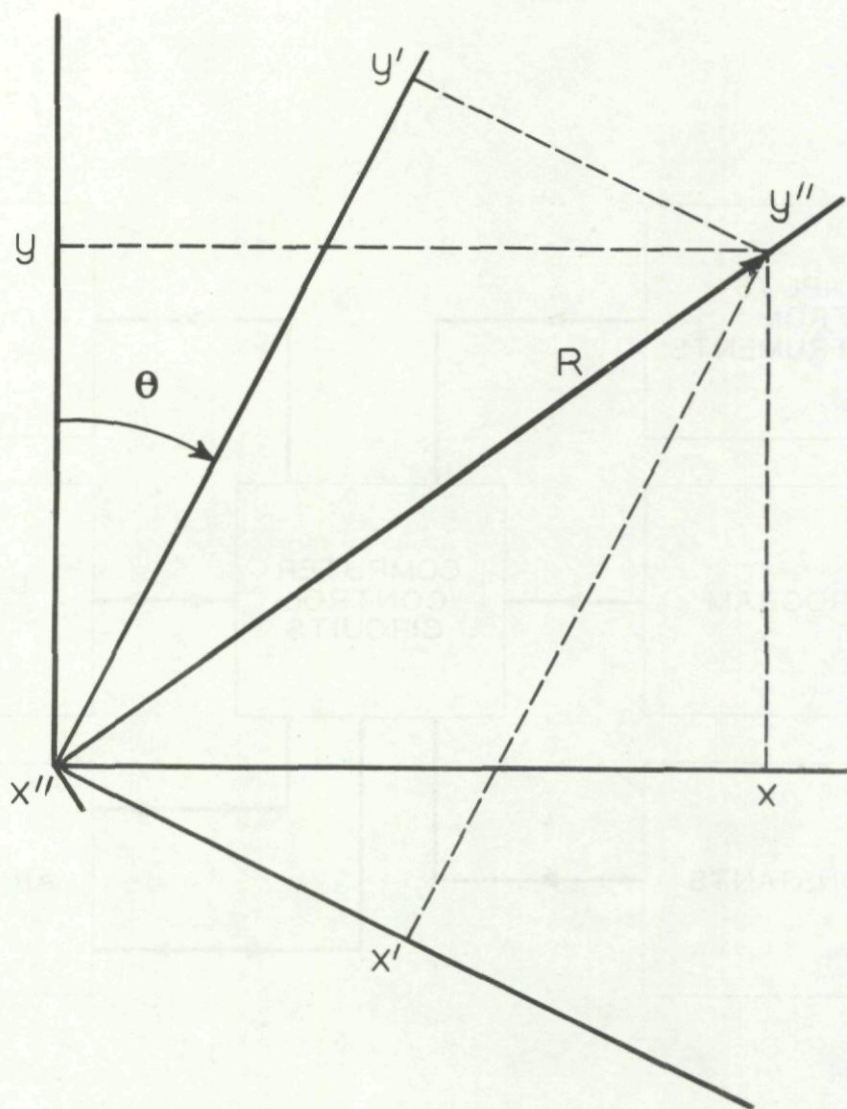


Fig. 6. Rotation of rectangular coordinates in two dimensions.

THE USE OF DIGITAL COMPUTER TECHNIQUES IN
MISSILE DESIGN AND CONTROL*
D. H. Gridley**

SUMMARY

The combined impact of high-speed electronic digital computers and the useful techniques derived from their development is now being felt in the areas of missile design and control. Two examples of digital processing are discussed as they apply to the pre-flight and postflight phases of missile system design and analysis. A third example is aimed at the control area with a presentation of applicable methods for the conversion between analog and digital quantities as required in the digital control system.

SOMMAIRE

L'utilisation combinée des calculatrices électroniques digitales à grande vitesse et des intéressantes techniques dérivées de leur développement est maintenant tombée dans le domaine de l'étude des missiles et de leur contrôle. Deux exemples d'application de procédés digitaux comme ils sont appliqués à l'étude et à l'analyse des phases d'avant et de post-vol d'un système de missile, sont discutés. Avec la présentation des méthodes applicables pour une conversion entre les quantités digitales et analogues comme cela est nécessaire en matière de système de contrôle digital, un troisième exemple est donné dans le domaine du contrôle.

1. INTRODUCTION

Over the past decade and a half the electronic digital computer has been brought out of the "giant-brain" category and placed into an extremely useful and practical area for use by both business and science. These machines have in turn been reduced in size, but increased in both speed and reliability. As a result of the growth period we have also learned considerable about the methods by which they can be used most efficiently. Another outgrowth of this era has been the

practical application of the digital technique to fields other than efficient computation.

These factors as they apply to the design and control of missiles will be the topics covered in this paper. An attempt will be made to point out three separate areas where the digital computer and its techniques are applicable - design computation, data handling, and transducers; yet your imagination will add others to this list which will be topics for future discussions. The first two areas are aimed at the design factors

*The opinions expressed in this paper are not necessarily those of the Department of Defense, the U. S. Navy, or the Naval Research Laboratory.

**Naval Research Laboratory, Washington, D. C.

of missile design, one on a preflight basis and the other on a postflight basis, whereas the latter factor is aimed at the control area.

The design-computation use of digital computers is probably the first area in which such machines were used for missile work. And, although we are still learning how to apply their capabilities, it appears that a plan for the organized use of their facilities is coming into being. At first the task was that of producing machines that would do the job. Here programming of problems was of secondary interest; but today, with machines available, the task of providing efficient problem programming has become formidable and is faced equally by all personnel using the machines. In particular fields, such as missile design, which may also include any vehicle design, there are certain basic sets of equations that describe a vehicle's general performance in relation to its configuration, balance, power, lift, atmospheric drag, and other external forces.

The task of setting up and checking out a program allowing for this multitude of variables is a lengthy job for several experienced mathematicians and programmers. However, it can probably be said that this same problem has been worked out in numerous groups faced with the need for a similar solution. It is here that coordinated effort by these groups can, and has, proven to be valuable.

In diverse groups, but where similar types of computers are available, an exchange of basic programs permits an efficient use of high grade personnel and relieves them of tedious duplication of programming tasks. It frees them for more detailed studies, perhaps even on areas described generally by the exchanged programs. This type of exchange need not be limited to groups possessing similar equipment, since interpretive routines have been worked for

changing between programs of single and triple address coding, and in many instances between very dissimilar machines. Nevertheless, continuity between computer orders, word lengths, and addressing methods should be considered by machine designers with this problem in mind.

It is felt that proven general purpose programs designed for flexibility of parameters, described either by single valued inputs or by variable values developed in branch subroutines, may prove to be as important to the missile designers as the wind tunnel has proven in the past. But organized and cooperative effort is required if progress is to be made.

An attempt at such a program is that describing a single stage rocket trajectory. Fig. 1 shows the basic factors concerned in the trajectory. Fig. 2 in turn shows the interdependence of the general program on the contributing parameters. Most of these can be considered in the program as variables with each being described, possibly by separate subroutines. It can be seen that the basic program set-up, assuming externally applied parameters, can be varied by altering the subroutines and their formulation. Although numerous difficulties may exist in the instrumentation of such an approach, the final results appear to offer many advantages and improved efficiencies in the use of computer facilities.

The foregoing discussion is a suggested pattern of attack aimed at efficient use of the computational capabilities of the digital computer which, if given early impetus, can be extremely useful as time goes on. The foregoing, however, is concerned primarily with efforts in the preflight design phase of missile vehicles or systems. The next item to be discussed is also related to design, but occurs in the postflight phase when everyone sits around and argues about why

the missile performed as it did or why it did not operate as expected. Postflight analysis encompasses the processing of numerous types of data; however, this discussion is restricted to the area of telemetered data.

2. DATA HANDLING

Until recently, telemetered data have been handled primarily on a basis whereby an analog strip record presentation has been made available for analysis by test and engineering personnel. Areas of interest can be quickly determined from such a presentation; nevertheless, detailed analysis of these areas required that they be read at frequent intervals, numerical values obtained, linearizing factors applied, and the data replotted before accurate quantitative analysis can be performed. Even with some machine aid this is a lengthy and tedious process continually encumbered with the human element. If the time schedule of a test program is compact (few are not), little or no time for this type of analysis can be allowed, and only the obvious failures are presented, but often without adequate reasons for their occurrence. Consequently, a rapid but completely automatic system is desired that will alleviate the problem for telemetered data reduction.

In the solution of this problem we have again made use of digital techniques. The desire to use the general purpose digital computer and its versatility is, at first glance, an obvious solution but after a further breakdown of the tasks to be accomplished, the problem appears too simple for such a machine and does not make efficient use of its capabilities. The equipment phase is further complicated by a desire that the reduction system be designed for field use and be trailer mounted.

A complete system design for the task of reducing any (fm/fm, pdm/fm, and ppm/am) telemetered data has resulted in an array of digital data reduction equipment that will best meet the field operation needs and still preserve the possibility of using the data obtained in a general purpose computer for the small amount of analysis not adequately provided for in the field unit. Data reduction times are reduced from months to that of hours and consistency of data is assured.

The first process for this equipment to perform is that of converting the telemetered analog quantities to digital data for subsequent recording in the digital form on multichannel magnetic tape. This operation is illustrated in Fig. 3. The input data are accepted in real-time, or played back in delayed-time from video tape recordings which were made in real-time. In either case the input data are sampled sequentially and directed to one of the two forms of quantizers - voltage, or pulse width. The ppm/am data is actually the distance between a zero reference pulse and an information pulse which is converted on the ground to pulse width data for quantization by the pulse width quantizer. The voltage quantizer can operate on input data sampled at rates up to 20 kc. The pulse width data quantizer operates at the input data rate of 900 or 5,000 samples per second.

We have chosen to quantize the data to 1 part in 256 or 8 binary places for ultimate data accuracies of $\pm 1/2$ percent, although all records are not equal to this precision. The eight bits of quantized data, sample by sample, are placed on the magnetic tape in parallel channels. Additional magnetic tape channels are reserved for coded-time values and frame synchronization pulses. Recording tape speeds are in five selectable steps, from 60 inches per second to 4-1/2 inches per second, one of which is selected for compatibility with the incoming data rate.

Once the data have been quantized and stored in digital form on the magnetic tape they are in a common format for use in the data reduction process or for ultimate insertion into a digital computer. An illustrative sample of these tape data is presented in Fig. 3 where 16 channels of ppm/am data are recorded at the 5 kc rate.

Before describing the reduction equipment, a look at the form of the usual output (Fig. 4) will illustrate better what must be accomplished. Shown is a section of visual fixed-styli type of output record having space for a total of 600 styli evenly spaced for 11 inches of the 12-inch paper width. 256 styli (or 512 styli, if every other stylus is used) are reserved for the basic data curve with all, or selected data points, printed linearly down the tape. Light vertical lines are written by every 10th or 20th stylus of the 256 scale, to provide rapid visual reference. A selection in the speed of the output record (continuously variable from 3/4 inch to 10 inches per second) can compress or expand the data into specified length scales for ease of reading.

To the left of the record, coded time markers are used to note each second of time from 0 to 9999 seconds or, if desired, in zone time, with hours, minutes, and seconds recorded to the nearest second. Light horizontal lines note the 1/10 second real-time intervals. To the right are 4-character Arabic numerals and a fiducial mark. These numerical characters are formed in a 5 x 7 dot pattern that can be written out as often as once every 1/10 second of real-time. The fiducial marker notes the exact curve reference point at which the numerical reading is made. The 4-character numerical data are carried by the system in order to permit "end-organ" reference readings to be portrayed. Both the curve data and the end-organ values are presented in the linearized form.

Fig. 5 portrays the relation between input and output binary or end-organ data, a conversion that is performed within the reduction portion of the system.

Reference to the reduction equipment can now be made (see Fig. 6). First is shown the magnetic tape playback equipment which has similar physical and tape transport characteristics as the recorder used previously. The digital record tapes are replayed at selected speeds but are under control of the Programmer Unit, which operates from coded timing track data. On the Programmer Unit control panel, selection can be made of the channel number to be transcribed, the time to start, the time to stop, transcription, and the recording rate (every data point, every other point, etc.) of data as the magnetic tape records.

Working in conjunction with the Programmer Unit is the Sub-Commutated Channel Selector, which is brought into use only when data of this type are used. It should be noted that the possibility exists that the selected output data may be quite low in reference to a real-time record; thus it is possible to speed up the playback tape unit and reduce a single channel at better than real-time.

Selected data passed by the Programmer Unit and/or Sub-Commutated Channel Selector are presented to the linearizer for alteration into the various forms to be used as output. The linearizer is a magnetic core matrix of 24 planes, each with 256 cores on a plane. The linearizer is set up with data obtained from a chart similar to Fig. 5, prior to playing back a specific channel. This set-up is performed by automatically reading a punched paper tape, prepared previously from the chart data. This read-in time is about 1 minute. At the output of the linearizer are the conversion equipments for producing the fixed-styli

recorder driving signals and for producing a punched paper tape (with format) in a form that can be used directly by a high-speed electronic digital computer (e.g., NRL's NAREC) if any further processing is desired.

The Arabic numeral generator is also a magnetic core device, with each character having a 5 x 7 core plane associated with it. One of ten possible character forming wires are energized from the linearizer signals, thereby setting up selected cores. The character is developed in proper form by synchronizing the matrix readout process with the chart speed. When all seven readout steps have been completed, the matrix has been cleared and a neutral array is available for receiving the next input signal.

3. CONVERSION PROCESSES

Digital techniques have not been used extensively within dynamic control loops primarily because adequate transducers have not been available; yet today, with the advent of reliable precision end instruments and miniaturized componentry, the many advantages and versatility of the variable programmed digital system has altered the thinking in many missile control groups. The progress in the area of producing operational equipment has been slow because of the overwhelming simplicity of analog instrumentation. Yet, with a view to the future, the digital technique is now developed to a point in theory, practice, and physical compatibility (size, weight, and power) that it should be considered more seriously as a part of the storehouse of techniques to be drawn upon when system designs are being developed.

Many control problems call for the solution of basically simple problems; however, we all know that within any of the confines of modern defensive vehicles such as aircraft,

ships, or missiles, the complexity of mathematical tasks to be performed makes the "computer-for-each-task" attitude obsolete. This observation comes not only from the many additional tasks that must be performed, but also from the realization that these tasks are composed of data selection and evaluation, both followed by computational operations requiring storage capacity and precision far beyond the scope of available or known analog techniques.

No single data input is now assumed desirable, but by selection of data from the many possible sources of intelligence, an evaluation is made and the data are used in the best combination for the situation presented at the moment. Basic data, as in the past, are the primary input to any computing system, but like the human system, the many contributing but normally nonessential data forms of information are necessary to make an automatic system function with the proper, "educated guess" in order to obtain the proper decision. The human element of command takes into account the basic factors of any situation; yet by knowledge of numerous other contributing factors the individual makes his decision, which is normally called "out guessing" the adversary. It is this added information and his ability to perform a logical mental reduction of all this information that makes the individual fit or not fit for command. Likewise, an automatic system must make optimum use of all sources of intelligence available to it.

Adversaries in the past made use of the fact that only the human could make such decisions of control and, as a result, have actually used humans expendably as control elements. In the modern world the speed and accuracy of decision required in control has in many instances outmoded the human. Consequently, the necessary element is a system, void of human deficiencies but capable of using only the necessary data, properly

processing it, and then making a logical and proper decision, thus causing the action to be carried out.

Digital techniques as presently developed do not offer the answer to all of these requirements, yet perhaps they do offer more answers than we have ever known in the past.

The unusual speeds of handling information by such devices and their high precision make them capable of rapid processing, and depending on the complication of the process and the required rate of data output necessary for the end use, it is often possible to operate these devices in a problem sharing mode. This fact explains why digital techniques are so appealing; one device can be used for a number of tasks, thereby eliminating the need for duplicate, weighty, and power and space consuming separate but distinct elements.

Although special purpose type computers using digital techniques are called for in such applications, the restrictive name applied here should not be confusing. Digital computers of special design need not be too different internally from the so-called general purpose type. However, only such equipment as that needed for the type of programming to be handled is designed into the special purpose equipment. The control provides basic arithmetic operations and the associated memory can provide the general aspect of the system control, thus permitting alterations of problem formulation when requirements are changed. This can be done without making the system obsolete.

Several analog-to-digital conversion devices have been discussed as they apply to the data reduction task. The voltage quantizer has at present a precision of about 0.05 percent but the pulse width quantizer, or variations of it, has provision limited primarily by the transducer used to generate

the pulse duration. Optically-read code wheels offer high precision, yet leave much to be desired in ruggedness for actual operational use. Also required of most operational conversion equipments is an output form capable of being carried through multiplexing or time sharing circuitry without loss of quality. For lower precision data, voltage switching can be handled by diode gates with a single quantizer used for all inputs. Fig. 7 shows how this system would appear.

For highly precise data this method falls short. One of the most readily handled and accurate methods of instrumenting this action is accomplished by using simple phase shifters such as conventional electrical resolvers. This process is illustrated in Fig. 8. The driving frequency for the phase shifter is generated from digital circuitry within the computer or if the basic computer is ground based, a minimum of airborne circuitry. Each base cycle of this frequency is produced by counting at some binary multiple frequency ($\times 256, 512, 1024$, etc.).

At the output of each phase-shifting transducer attached to a control element will appear a new signal of identical frequency but phase shifted in an amount proportional to the input shaft displacement. Quantization of this phase difference by the digital system provides input for further digital processing. By selecting high driving frequencies (400 or 1000 cps) the phase lag due to high input velocities and acceleration is minimized. For higher precision operation two- or three-speed systems may be required. In these instances the multiple transducer units are geared together in ratios related to the number system used within the computer. Ambiguity between digit readings, derived from the several transducers in a multi-speed system is resolved logically within the digital system.

The computer output (Fig. 9) can be either a binary-coded group of signals that are converted into a pulse position relative to the base frequency zero point or it can generate the pulse position internally. In either instance the external end instrument drives to match the pulse positions from the computer and that of zero crossover point of the transducer output signal. The advantage of this method of conversion is that in both the analog-to-digital and digital-to-analog conversion process, the information is carried in a pulse form which can be handled to a high degree of precision by simple logical diode gating arrays.

Only general methods have been described here and no attempt has been made to discuss the mathematical theory of sampled data servo loops. Yet, in general, it is pointed out that dynamic equalization can be provided either within the computer or by analog corrective networks within the external feedback loop. In all probability both could be used if such complications are required; but with the computer providing data composed of gross values and nonlinear function, analog networks would handle the more conventional linear modifiers relating to the usual physical dynamics of a vehicle. Consequently, we have at our disposal a

tool for mixing the common with the uncommon data. It should not be overlooked that a digital system of this type also affords means for incorporation of a reasonably high precision data transmission system between missile and ground or fixed base elements. By use of these techniques a coded signal can be received or processed for transmission. Security of these signals is also improved by means of code redundancy favorably used for checking and error correcting.

Many control functions, navigation, or programmed slow maneuvers need not impose the high-frequency capabilities upon the computer system and can readily be handled at a much slower problem solution rate by the computer but interlaced with other, possibly unrelated solutions.

In summary, it is pointed out that the physical instrumentation and the ability to apply the digital techniques are progressing in a parallel fashion. The future holds much promise for a more complete application of their capabilities. Yet, a word of caution must be added. Digital systems are not the answer to or a panacea for all problems; consequently, a thorough system design, and evaluation of proposed applications should be reviewed with an open mind and selection should be based on overall merit.

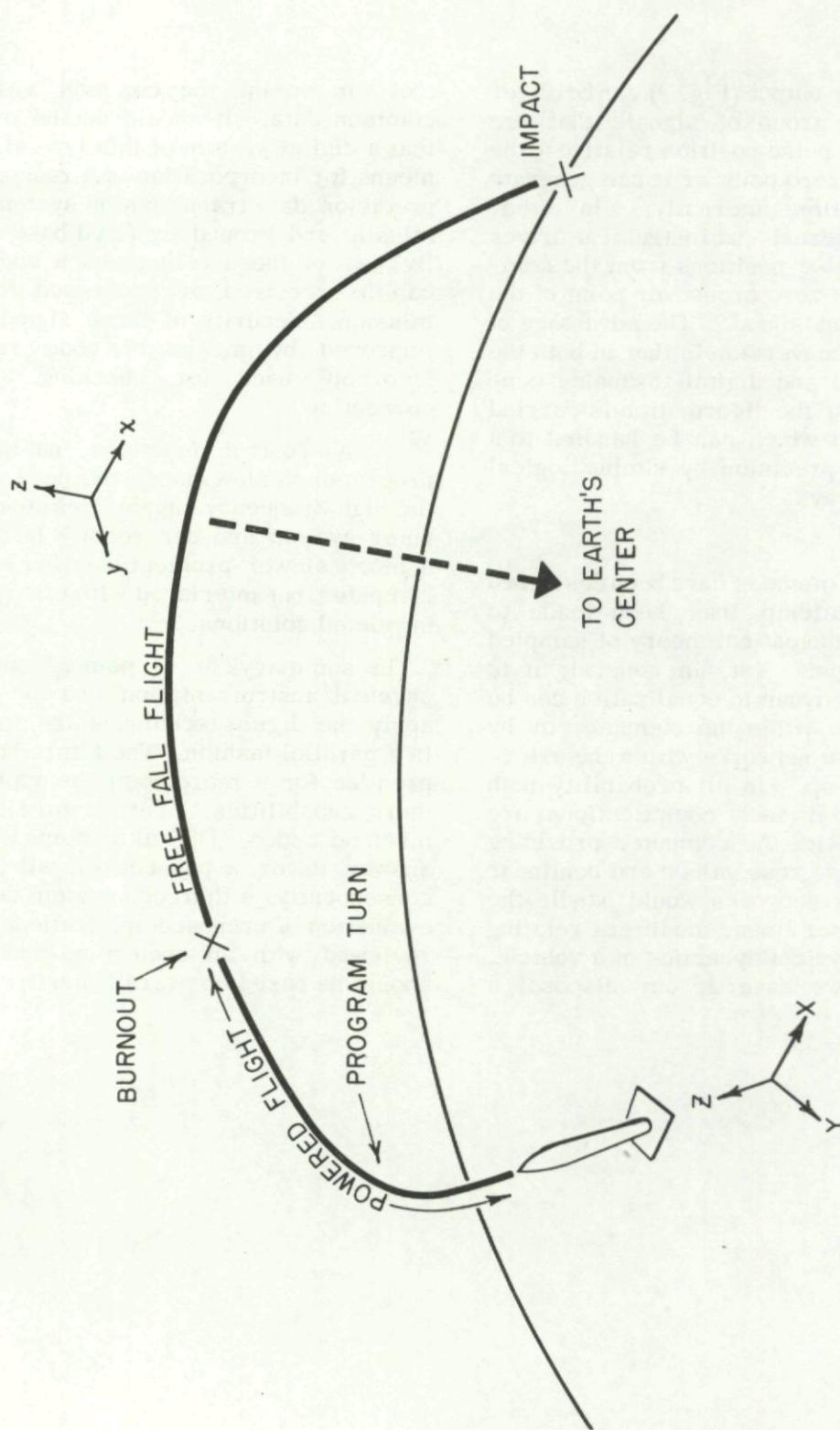


Fig. 1. Single-stage rocket, sample trajectory.

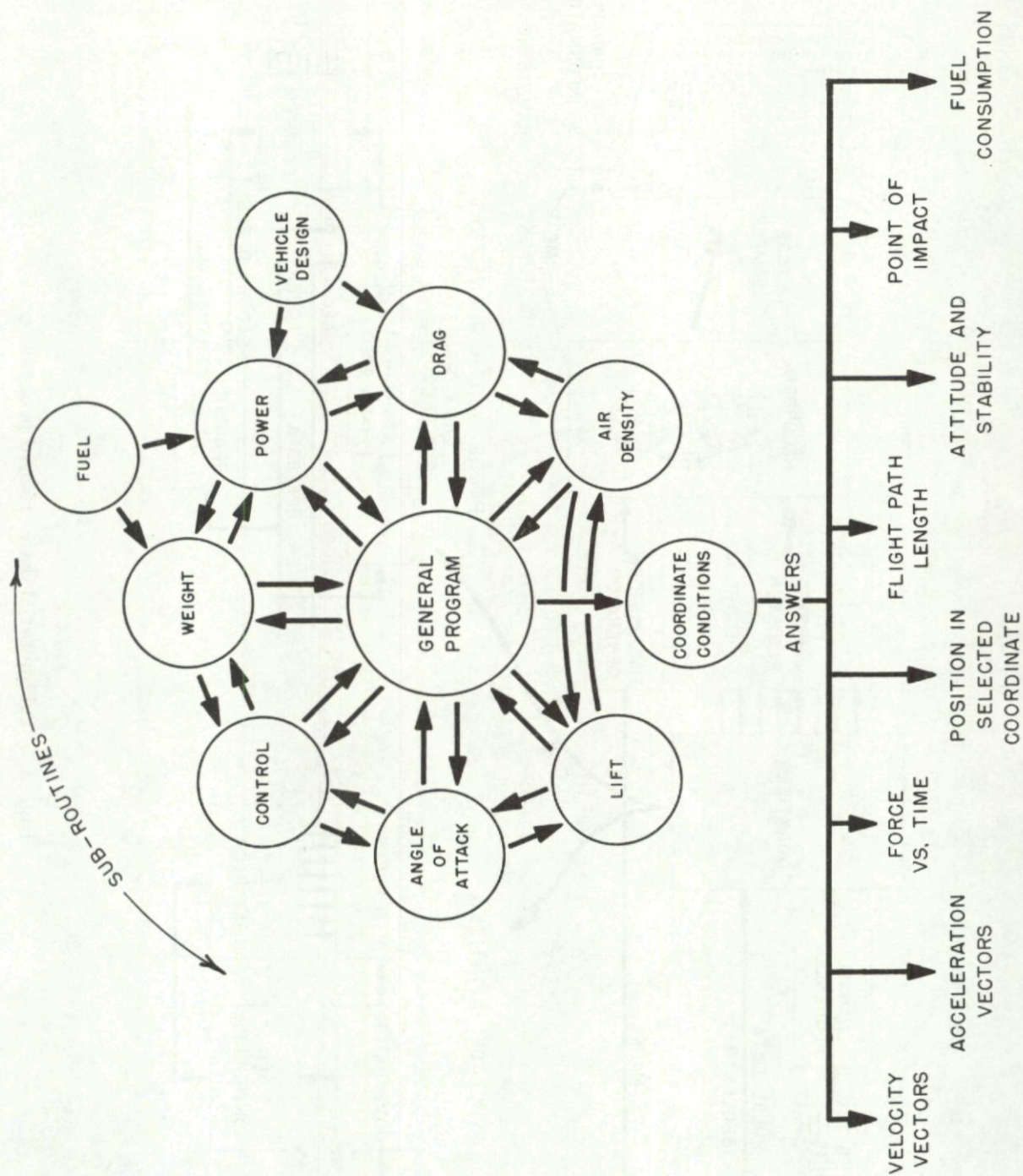


Fig. 2. The use of a general computer program.

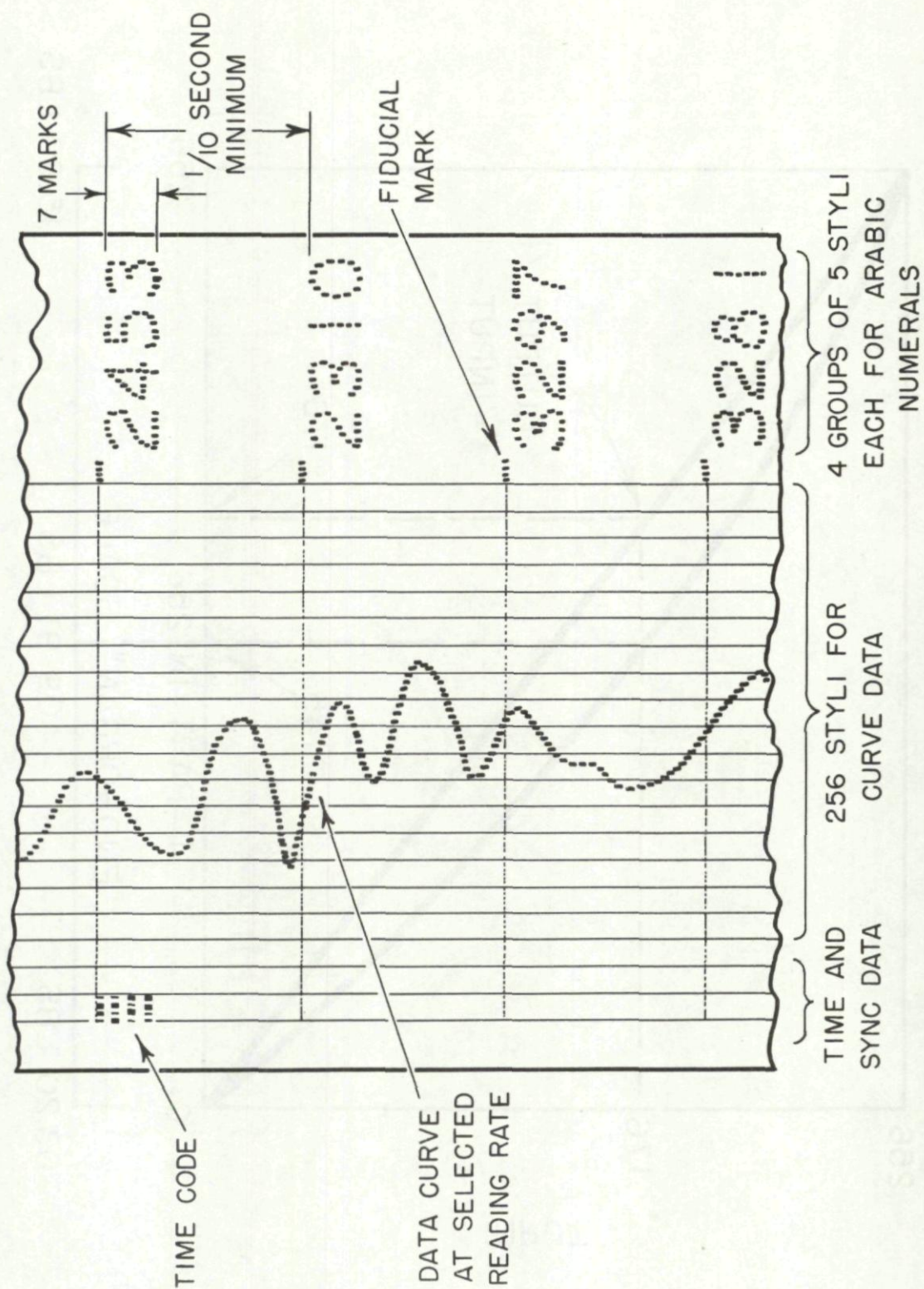


Fig. 4. Fixed styli record (sample).

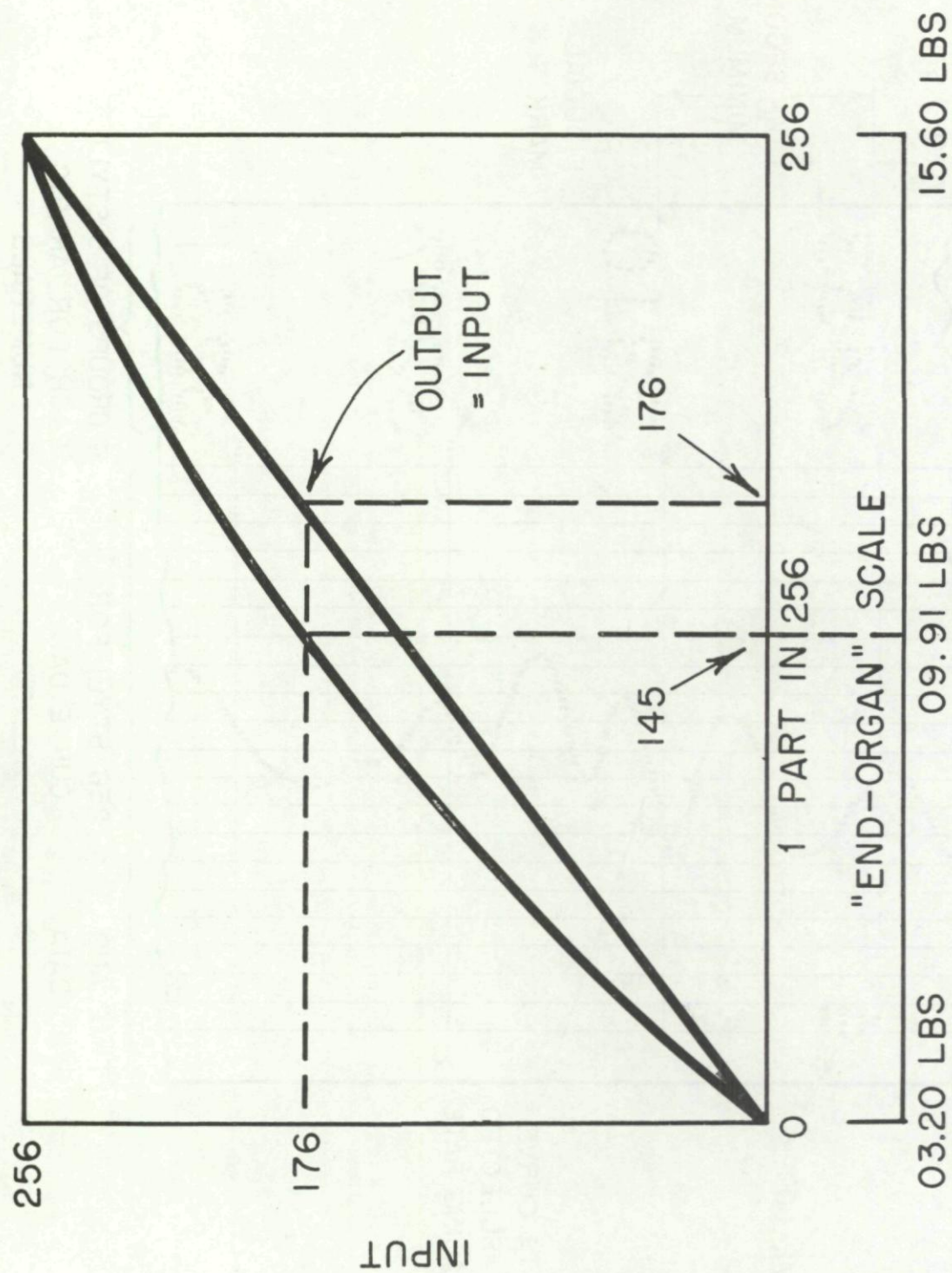


Fig. 5. Linearization chart (sample).

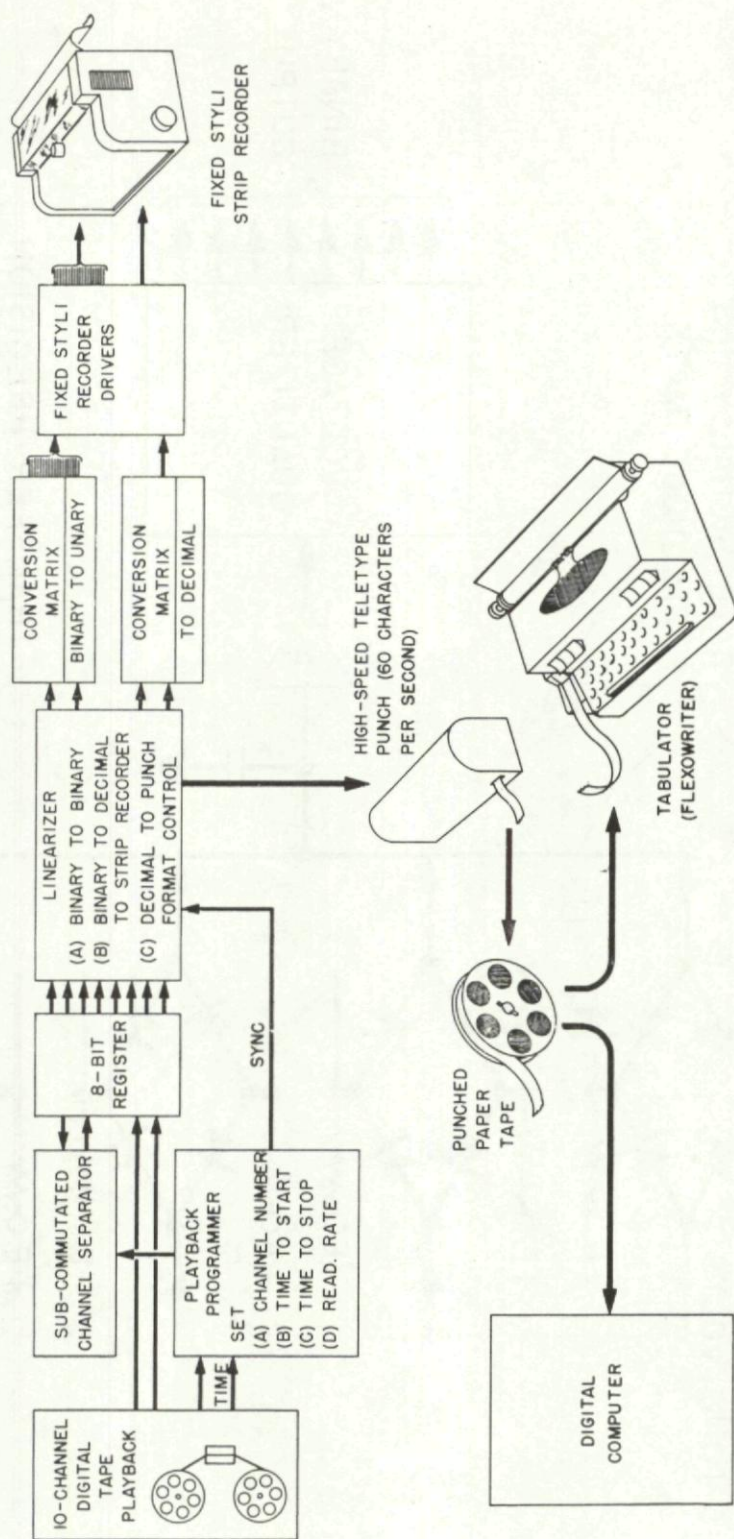


Fig. 6. Data reduction process.

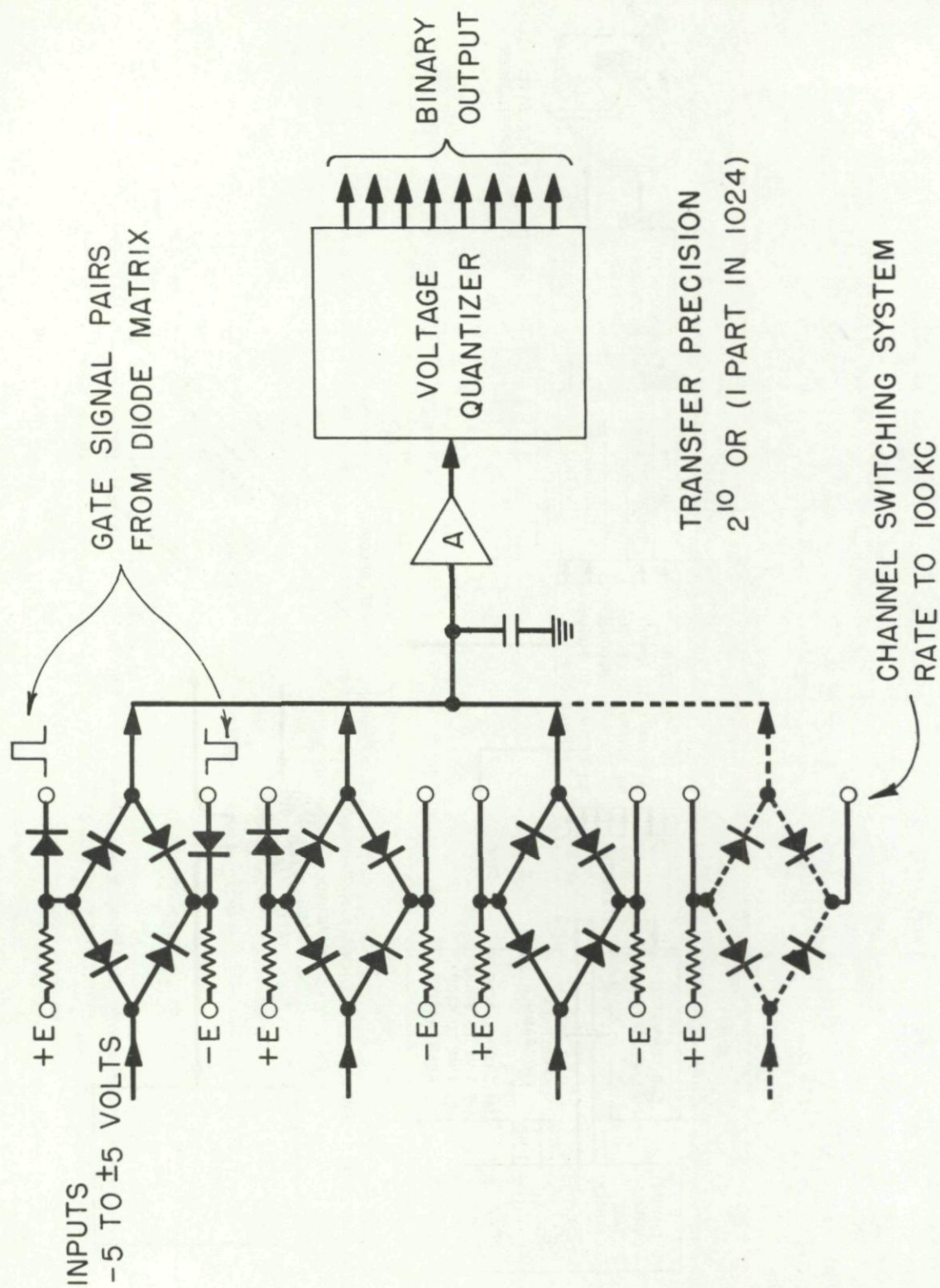


Fig. 7. Voltage sampler and quantizer.

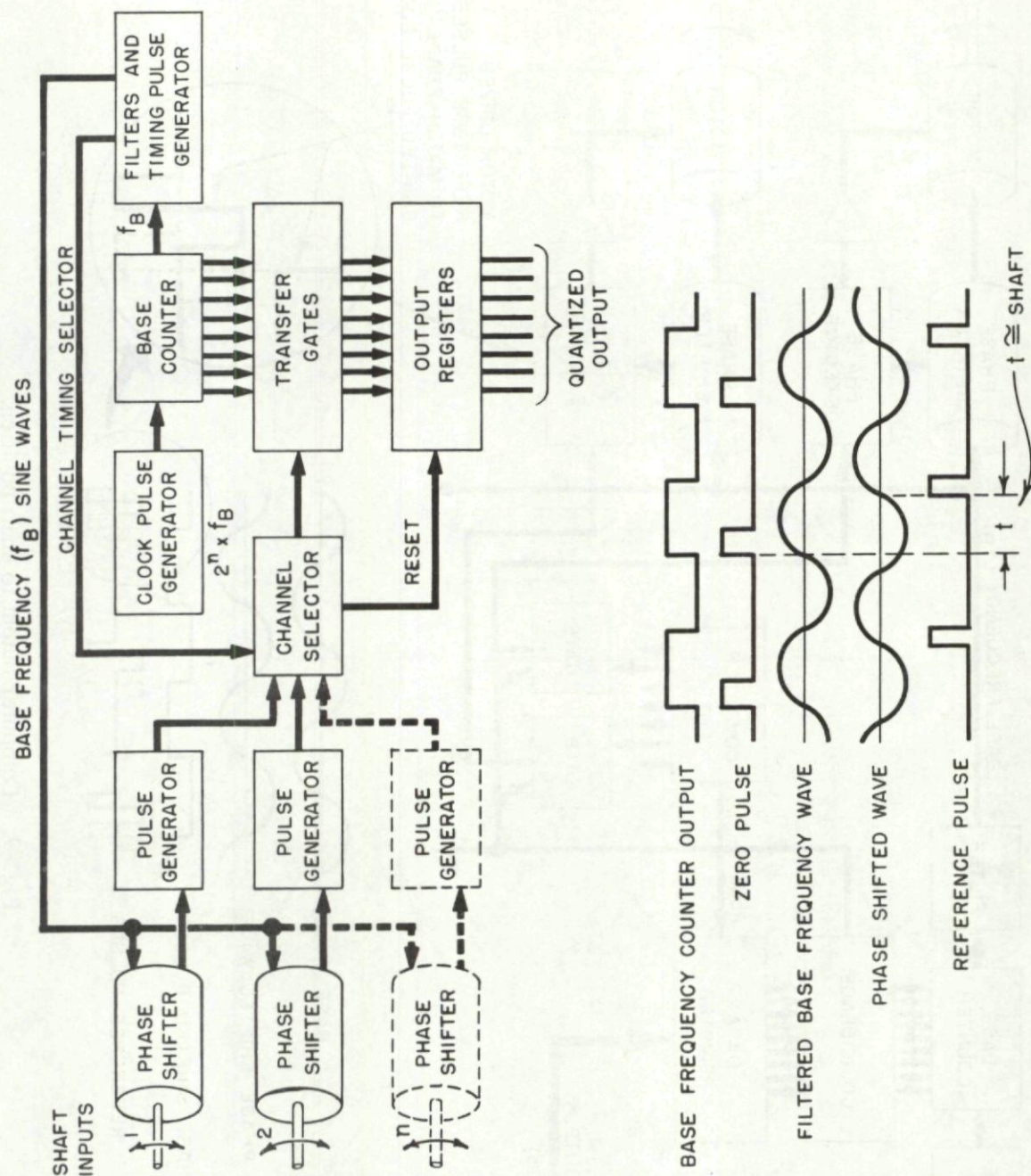


Fig. 8. Shaft position to digits.



Fig. 9. Computer output to shaft position.

THE APPLICATION OF NOISE AND FILTER
THEORIES TO GUIDANCE PROBLEMS
R. J. Parks and Robert M. Stewart*

SUMMARY

This paper presents the application of recent noise and filter theories to the design of guidance systems. It includes the application of Wiener single and multiple input filter theory to guidance system design, particularly where there is more than one source of guidance information with different characteristics and also to the case where one or more of the sources of information is received in the form of sampled data. Typical examples of the application of these techniques to practical problems are given.

SOMMAIRE

Cette note présente l'application des théories récentes relatives aux bruits et au filtrage à l'étude des systèmes de gouverne. Elle comprend l'application de la théorie du filtre d'entrée unique ou multiple de WIENER à l'étude des systèmes de gouverne, en particulier dans le cas d'une source d'information de gouverne avec différentes caractéristiques et aussi dans le cas où une ou plusieurs des sources d'information sont reçues sur la forme de données pulsatoires. Des exemples typiques de l'application de ces techniques à des problèmes pratiques sont donnés.

1. INTRODUCTION

This paper discusses the application of some of the recent noise and filter theories to the design of certain types of guidance systems. The command type guidance system for certain applications would, in general, be the one in which these concepts have the most direct application, but they might well find application in direct or modified form to other types of guidance systems, or in fact to many other related fields.

2. THE DISTORTIONLESS FILTER PROBLEM

Inherent in any guidance system is the measurement of one or more physical parameters such as position or velocity. It

is in general required that these measurements be quite accurate and also free of any significant filter lags or time delays. These two requirements are often conflicting inasmuch as the high accuracy requirement often dictates considerable filtering of the data which, of course, will in general cause some lag, delay, or distortion of the data. This problem has led to the concept of quasi-distortionless filter design in which, for certain forms of signals (for instance, signals which could be expanded in a power series in time with a limited number of terms) a filter can be designed that will cause no distortion. If the actual signal to be filtered deviates at all from the form assumed, the filter will distort the signal to a degree which obviously depends upon the magnitude of the deviation from the assumed

*Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California.

form. It depends upon the application as to how well the actual signals can be approximated by the forms which would allow the required amount of quasi-distortionless filtering, but cases do exist where there appears to be very little theoretical justification for this approach.

If the physical parameter of interest can be measured by two techniques which have significant differences in the characteristics of these error terms, particularly in relation to the spectral distributions of the error terms, a filtering technique is possible which will permit complete lack of distortion of the signal no matter what its form might be.

For instance, velocity can be measured by differentiating a position measurement and by integrating an acceleration measurement. The error characteristics of these two measurements would be considerably different. The first signal, because of the differentiation, would tend to have much relatively high-frequency noise and little error close to zero frequency while the second, because of the integration, would tend to have an error of just the reverse characteristics with most of its error being of the nature of a low frequency drift.

Consider, then, the diagram shown in Fig. 1. F_a and F_r are filter transfer functions for the accelerometer and position data, respectively; N_a is the error or noise of the accelerometer; N_r is the error or noise in the differential position data and V^* is the estimate of V made by this circuit.

It can be easily shown by analysis of the circuit in Fig. 1 that the error in the estimation is given by the transform or operational expression

$$\epsilon_v = V^* - V = [F_r + p F_a - 1] V + F_r N_r + F_a N_a. \quad (1)$$

It is to be noted that, if $F_r + p F_a = 1$, the system will be distortionless; i.e., the first term, which represents an error that is a function of the time history of the true velocity, becomes zero. Under this condition, the choice of F_a and F_r and the error in estimating V do not depend upon V itself but only upon N_a and N_r ; also, the error will have zero mean if N_a and N_r have zero mean, and the error will tend to zero if N_a and N_r approach zero.

If V had zero mean and resulted from a stationary random process and it was desired to minimize the time average of the squared error, the distortionless restriction would, in general, make it impossible to obtain as small an error as would otherwise be possible. Even in this instance, however, the difference between the absolute optimum and the distortionless optimum is negligible if the error in the accelerometer is small compared with the actual acceleration. Further, in many cases, the following factors apply:

- (1) V does not in general have zero mean and is not representative of a stationary random process.
- (2) In many instances, the value of interest may be the ensemble average of the error (as opposed to the time average) shortly after a near discontinuity in V .
- (3) Using the distortionless restriction, a decrease in the noise would result in a larger percentage drop in the error, with a lower limit of zero error.

For these reasons, it is believed that the distortionless criterion is a useful and valuable one.

It will now be assumed that N_r and N_a have zero mean and are stationary and random. With the possible exception of the stationary restriction, these assumptions are generally realistic for N_r , the error in the differentiated position data. Even the stationary assumption is not unrealistic since it is actually required only that N_r be approximately stationary over the weighting period of F_r . As a general rule, N_a will not have zero mean; however, the DC shift or mean-value error can be analyzed separately and by controlling the actual level of the zero shift not permitted to have a significant contribution.

These assumptions and the distortionless criteria result in the spectral density Φ_v of the velocity error $\epsilon_v = V^* - V$, being given by the expression

$$\Phi_v = |F_r|^2 \Phi_r + |1 - F_r|^2 \frac{\Phi_a}{\omega^2} \quad (2)$$

where Φ_r and Φ_a are the spectral densities of N_r and N_a , respectively, and $\omega = 2\pi f$, where f is the frequency.

Since the mean-square error and the error spectral density are related by

$$\sigma_v^2 = \overline{\epsilon_v^2} = \int_{-\infty}^{\infty} \Phi_v df \quad (3)$$

it is possible to find the form of F_r that will minimize σ_v if Φ_a and Φ_r are known.

The process of calculating the best form of F_r is given by Norbert Wiener (Ref. 1). The form of the equation for the error spectral density is the same as the equation (in the frequency domain) solved by Wiener for a single input channel.

A significant parameter of interest would be the accuracy improvement afforded by the combination measurement over an optimized system using only the position data. This comparison is difficult, however, since a position measurement alone with any filtering at all cannot be made distortionless, and the improvement afforded by the distortionless feature depends heavily on the V program of the particular system under consideration. However, in principle, it would be possible to program the standard values (ensemble average) of acceleration and to use these signals in lieu of the acceleration measurement.

If this programming were done, the effective acceleration error would be the difference between the actual acceleration and the standard acceleration. By definition, this error would have zero mean in the ensemble sense, and if, in order to gain a qualitative feeling for the improvement, it is further assumed that the error is random and stationary and has a flat spectral density Φ_T over the frequencies of interest, and that the spectral densities of the position error and acceleration error are also flat, it can be shown that the ratio of the velocity error for the combination measurement to that of the all-radio measurement is given by the expression

$$\frac{[\sigma_v]_{\text{comb}}}{[\sigma_v]_{\text{radio}}} = \left(\frac{\Phi_a}{\Phi_T} \right)^{\frac{3}{8}} \quad (4)$$

In general, then, if the actual acceleration deviated by about 10 percent from its standard value, the use of an accelerometer accurate within 1 percent would result in a factor of about 5.6 improvement in the velocity accuracy, and the use of an accelerometer accurate within 0.1 percent would result in a factor of about 30 improvement in velocity accuracy.

In practice, it would often be impractical actually to program the standard value of acceleration, unless it could be approximated by a constant over the region of interest. If the standard acceleration is not programmed, or if it is only approximated, and/or if the acceleration deviation from standard is not stationary (as it will not be in certain applications where the measurement is to be taken shortly after a near discontinuity in acceleration), the improvement afforded by the combination measurement will actually be underestimated by Eq. (4).

The improvement afforded by the combination filtering is very definitely a function of the degree of difference between the spectral density of the two measurement systems, and the above results apply only if the assumptions made apply. Obviously, if the two systems had identical error spectral densities, the combination estimate error could be, at most, a factor of the $\sqrt{2}$ less than the estimate that could be obtained from either measurement alone. Even this improvement would be attained only if the two measurement errors were independent.

Although velocity has been used as an example here, the same process could be applied to the measurement of any physical parameter that could be measured in more than one way with sufficiently different error behavior. Further there is no limitation on the number of measurements which could be combined. If three or more measurements were made of the same parameter, they could all be combined simultaneously to result in an estimation whose accuracy would be better than could be obtained if any one or several were not used. The degree to which any one measurement would improve the estimate would again depend heavily on how different its error spectral density were from that of any of the other measurements and, of course, on the relative magnitude of the error of that source.

3. THE APPLICATION OF WIENER'S RESULTS TO CLOSED LOOP SERVOS OR GUIDANCE SYSTEMS

The basic problem covered by N. Wiener in Ref. 1 is, in simplified form, essentially as follows. Consider a series of measurements of a signal all contaminated by noise (in general, different noise in each channel) in which each signal will be independently filtered and then all added together as shown in Fig. 2. What is the form of each of the filters Y_j that will result in a minimum r.m.s. difference between the output and the true signal? The following assumptions are made in Wiener's solution to the problem:

- (1) The filters are all linear.
- (2) The signal and the various noises are all stationary random series with zero mean and have known correlation functions or spectral densities.

Wiener's book presents a procedure for solving for the various Y_j . Consider now the feedback system represented by Fig. 3.

It can be shown easily that, if Y_2 is known and it is desired to determine Y_1 so as to minimize the r.m.s. difference between e_0 and S , the error can be represented in the form

$$e_0 - S = \frac{1}{1 + Y_1 Y_2} (Y_2 N_2 - S) + \frac{Y_1 Y_2}{1 + Y_1 Y_2} N_1 \quad (5)$$

which, if we let $Y = (Y_1 Y_2)/(1 + Y_1 Y_2)$ will be of the same form as the single input Wiener problem where Y is the unknown, $Y_2 N_2 - S$ is the equivalent signal, N_1 is the equivalent noise and, as such, can be solved by the technique outlined by Wiener. Once Y is determined, Y_1 can be easily determined, since Y_2 is known.

Consider the problem of guiding a drone airplane down a straight line by means of a ground based measurement of the drone error in position and through the use of a command link to the drone to cause it to maneuver to remove any error.

The system shown in Fig. 3, then, can be interpreted to cover this problem. The signal S is identically zero since it is desired to minimize the side displacement (represented by e_0) under the effects of both the noise or error in the ground measurements N_1 and the disturbances acting on the drone represented by N_2 . The disturbances in general would be caused primarily by side wind effects and inaccuracies of the drone autopilot.

The transfer function Y_2 would represent the effects of these disturbances on the actual side displacement. Since these effects would be determined by the airframe and propulsion unit of the drone, they could be known for any particular drone. The unknown Y_1 would represent the filtering that should be done on the measured error signal before it is transmitted to the drone and mixed into the drone autopilot input. If more than one disturbance source exists, they can all be represented by an equivalent wind disturbance, for instance.

For example, if it is assumed that

$$Y_2 = \frac{1}{p(1 + Tp)} ,$$

which would be representative of a wind disturbance in which a sudden wind would first cause a side acceleration which would gradually decay until the drone took up the side velocity of the wind, and that N_1 and N_2 both had flat spectral density of magnitude ϵ^2 and 1 respectively (assuming a unit spectral density for ϵ^2 causes no loss

of generality since it merely defines the units being used in the problem), the solution for Y_1 is

$$Y_1 = \frac{1 + (\sqrt{\epsilon^2 + 2\epsilon T} - \epsilon)p}{\epsilon} .$$

This solution indicates that no filtering of the ground signal should be done even though it is noisy; in fact, the ground signal should have a derivative term added which will make it even more noisy. The gain of the ground operation is specified as $1/\epsilon$ and is therefore a function of the ground noise. The result indicates that the filtering is really done by the drone airframe itself. Even though the command signal transmitted to the drone is extremely noisy, the actual side position of the drone will be very little affected by the noise due to the fact that the side position of the drone cannot be changed rapidly and it therefore does not respond to or filter out the noise in the command signal.

This system would, in fact, be impractical as it stands, since the r.m.s. value of the noise transmitted over the command channel would, in the example, be infinite. This situation is due to the fact that the noise spectral densities were assumed to be flat out to infinite frequencies. In practice, they could be filtered with filters with time constants small compared to the rest of the system so as to have a bandwidth large compared to the bandwidth of the overall servo loop but still not be infinite. Furthermore, there exists a straightforward extension of the technique used which permits the optimization of the error signal in the same sense as before but subject to the restriction that the r.m.s. value of the noise in the command channel or any other point in the circuit be equal to or less than some specified value. Obviously, the more the

r.m.s. command noise level is restricted, the more the restricted optimum will be worse than the unrestricted optimum. In general, however, a compromise can be found that results in an acceptable command noise level without seriously reducing the overall guidance accuracy.

Suppose further that the drone had aboard a gyro combined with a type of drift measurement or had aboard a simple form of an inertial guidance system that would permit some form of crude but not sufficiently accurate guidance even in the absence of the ground command system. The question then naturally arises of how much of each guidance signal should be used and of how the signals should be mixed to permit the minimum guidance error. This problem can be handled by an extension of the above technique which is parallel to Wiener dual input problem.

This problem can be represented diagrammatically by Fig. 4, where Y_{22} is known and represents the airframe response, N_3 represents the disturbances affecting the drone, and N_1 and N_2 represent the noise or error of the two guidance measurements respectively. The transfer functions Y_{11} and Y_{12} are the unknown, and it is desired to choose them so as to minimize the r.m.s. value of the difference between e_0 and S .

In this particular case, S is again identically zero, since, in this example, it is desired to guide the drone down a straight line (if the desired path were something other than a straight line, the signal S would represent the difference between the actual desired path and a straight line).

It can be easily shown that this problem can be reduced to the problem shown in Fig. 5, where it is desired to choose Y_1 and

Y_2 so as to minimize the r.m.s. difference between S^* and S , if

$$Y_1 = \frac{Y_{11} Y_{22}}{1 + (Y_{11} + Y_{12}) Y_{22}}$$

and

$$Y_2 = \frac{Y_{12} Y_{22}}{1 + (Y_{11} + Y_{12}) Y_{22}}$$

(6)

and $Y_{22} N_2$ is the equivalent signal.

The problem shown in Fig. 5 is, of course, the special dual input case of the general problem which was shown in Fig. 2 and solved by Wiener (see Ref. 1).

Following the procedures given in Ref. 1, it is possible to solve for Y_1 and Y_2 and therefore for Y_{11} and Y_{12} from Eq. (6). It is further possible, if the particular problem warrants it, to obtain a conditioned optimum solution to this problem subject to the limitation that the r.m.s. error at some point in the circuit, say at the command channel point, be no greater than some specified amount.

This particular problem can be further extended to include any number of inputs; it is not limited to one or two. Unfortunately, except in some special cases, the mechanics of solving the problem become more difficult whenever two or more inputs are required, but the procedures are outlined and can, in any specific case, be carried through with some labor.

4. AN APPLICATION OF WIENER'S MULTIPLE-INPUT FILTERING THEORY TO DIVERSITY SYSTEMS

A technique of considerable interest in contemporary communication, control, and instrumentation systems is illustrated in Fig. 6.

Wiener's general analysis of such problems shows that, if S, N_1, N_2, \dots, N_n are mutually independent stationary random time-series, the impulsive-responses $y_i(\tau)$ of the linear filters which minimize the mean-square error in estimating the "signal" S by S^* satisfy the following simultaneous set of integral equations: ($j = 1, 2, \dots, n$)

$$\begin{aligned} \phi_s(\tau) = \int_0^{\infty} \left\{ \left[\phi_s(\tau - \sigma) \right] \sum_{i=1}^n Y_i(\sigma) \right. \\ \left. + \left[\phi_{N_j}(\tau - \sigma) \right] Y_j(\sigma) \right\} d\sigma \end{aligned} \quad (7)$$

for $\tau > 0$, where the ϕ terms are autocorrelation functions of the signal and various noises.

These equations are equivalent to the set

$$\begin{aligned} \phi_s(\tau) - \int_{-\infty}^{+\infty} \left\{ \left[\phi_s(\tau - \sigma) \right] \sum_{i=1}^n Y_i(\sigma) \right. \\ \left. + \left[\phi_{N_j}(\tau - \sigma) \right] Y_j(\sigma) \right\} d\sigma = h(\tau) \end{aligned} \quad (8)$$

with the understanding that

$$Y_i(\sigma) = 0 \quad (9)$$

for $\sigma < 0$, (the "realizability" condition) and

$$h(\tau) = 0 \quad (10)$$

for $\tau > 0$.

Fourier-transforming both sides of Eq. (8) then gives:

$$\Phi_s(\omega) \left[1 - \sum_{i=1}^n Y_i(\omega) \right] - \Phi_{N_j}(\omega) Y_j(\omega) = H_j(\omega) \quad (11)$$

where the Φ 's are spectral densities of signal and the various noises, the Y 's are frequency response functions of the optimum filters having (from Eq. (9)) no poles in the lower half-plane, and the H 's having (from Eq. (10)) no poles in the upper half-plane.

A frequently encountered situation is that in which the noises in the various channels are statistically similar, except possibly for amplitude. Then,

$$\Phi_{N_j} = K_j^2 \Phi_{N_0} \quad (12)$$

Substituting Eq. (12) into Eq. (11) then gives:

$$\Phi_s \left[1 - \sum_{i=1}^n Y_i \right] - K_j^2 \Phi_{N_0} Y_j = H_j \quad (13)$$

Subtracting each side of this equation for arbitrary j from that for $j = 1$, e.g., gives

$$\Phi_{N_0} [K_j^2 Y_j - K_1^2 Y_1] = H_1 - H_j \quad (14)$$

Using Wiener's spectral-factorization theorem,

$$\Phi_{N_0}(\omega) = \psi(\omega) \overline{\psi(\bar{\omega})} = \psi^+ \psi^- \quad (15)$$

where ψ^+ and ψ^- have poles and zeroes only in the upper and lower half-planes respectively. Thus,

$$\psi^+ [K_j^2 Y_j - K_1^2 Y_1] = \frac{H_1 - H_j}{\psi^-} \quad (16)$$

The left side of Eq. (16) can then have poles only in the upper half-plane, while the right side has poles only in the lower half-plane; hence,

$$\psi^* [K_j^2 Y_j - K_i^2 Y_i] = \text{constant}. \quad (17)$$

It may be easily shown that as $\omega \rightarrow \infty$,

$$\psi^* [K_j^2 Y_j - K_i^2 Y_i] \rightarrow 0 \quad (18)$$

and thus, in general

$$K_j^2 Y_j - K_i^2 Y_i = 0 \quad (19)$$

or

$$Y_j(\omega) = \frac{K_i^2}{K_j^2} Y_i(\omega). \quad (20)$$

Eq. (20) indicates that the optimum system may be mechanized with just one frequency-sensitive element as shown in Fig. 7. This implies the somewhat surprising result that the optimum bandwidth for each channel is identical, even though the signal-to-noise ratios may be quite different.

The best form of Y_i may be determined as follows: Substitution of Eq. (20) into Eq. (13) gives

$$\Phi_s \left[1 - K_i^2 Y_i \sum_{j=1}^n \frac{1}{K_j^2} \right] - K_i^2 \Phi_{N_0} Y_i = H_i \quad (21)$$

or,

$$\Phi_s - \left[\Phi_s + \Phi_{N_0}^i \right] Y_i^i = H_i \quad (22)$$

where

$$\Phi_{N_0}^i = \frac{\Phi_{N_0}}{\sum_{j=1}^n \frac{1}{K_j^2}} \quad (23)$$

and

$$Y_i^i = \left[K_i^2 \sum_{j=1}^n \frac{1}{K_j^2} \right] Y_i. \quad (24)$$

Eq. (22) may be solved for Y_i^i by the method applicable to a single-input filter, and then Y_i may be found from Eq. (24).

5. SOME NEW RESULTS CONCERNING THE STATISTICAL DESIGN AND ANALYSIS OF SAMPLED-DATA SYSTEMS

a. Restoration of Sampled Data

Ref. 2 describes a statistical technique applicable to the problem of restoring sampled data. In this section we will merely summarize the principal results obtained.

The problem treated can be represented by the block diagram in Fig. 8. According to the well-known sampling theorem, if $m(t)$ has no frequency components higher than f_0 , all of $m(t)$ can be perfectly reconstructed from samples taken periodically at any frequency greater than $2f_0$. Since this is an idealization of the usual true situation, it is of interest to examine the relationship between sampling frequency and minimum attainable errors in the output for messages having arbitrary spectra.

Using mean-square error as a measure of fidelity, the following major results are obtained for the case of noise-free sample pulses:

(1) A straightforward procedure is developed for determining the optimum physically realizable time-invariant linear smoothing filter. It is shown that the error spectral density is given by

$$\Phi_e(f) = |Y^1 - 1|^2 \Phi_m(f) + |Y^1|^2 \Phi_s(f) \quad (25)$$

where

$$Y^1 = f_s Y \quad (26)$$

and the sideband spectrum Φ_s is

$$\Phi_s(f) = \sum_{n=1}^{\infty} [\Phi_m(f - nf_s) + \Phi_m(f + nf_s)] \quad (27)$$

This spectrum is illustrated in Fig. 9. It may be noted that Eq. (25) is identical in form with the well-known expression for error spectral density for continuous smoothing of message plus noise if the noise were independent of the message and had a spectral density equal to the sideband spectrum Eq. (27).

Hence Wiener's method may be applied directly to give the optimum (in the sense of minimum r.m.s. error) linear filter function Y^1 (and $Y = Y^1/f_s$) for any given sampling rate f_s . Eq. (25) can be easily modified to allow for delay in recovering the message if such delay is desirable for greater accuracy.

(2) There is no time-varying linear filter (whose characteristics vary periodically in synchronism with the sampling frequency) which is better than the optimum time-invariant linear filter.

(3) The weighting function $y(\tau)$ of such an optimum filter should always be equal to unity for $\tau = D$, where D is the desired delay in recovery of the message,

and should equal zero for $\tau = D + n\Delta$, where n is any integer other than zero, and Δ is the sampling interval. For physical realizability, $y(\tau) \equiv 0$ for $\tau < 0$ (see Fig. 10).

(4) If the spectral density of the original data or message is of the form $K/(f_0^2 + f^2)$ and if these optimum recovery filters are used, the sampling frequency must be of the order of $10^4 f_0$ in order to obtain 1 percent error in the reproduction.

(5) If the spectral density of the original message at frequencies near half the sampling frequency and above is approximately of the form K/f^{2n} the r.m.s. error σ_e varies with sampling frequency f_s approximately as $f_s^{-(n-1/2)}$.

b. Sampled Data in Closed-Loop Systems

Consider the closed-loop control system indicated by Fig. 11.

By methods similar to those of Ref. 2, it is easily shown that the Fourier transforms of error ϵ and disturbance D over a long but finite time $-T \rightarrow +T$ are related approximately* by

$$\epsilon_T(f) = Y_m(f) [D_T(f) - f_s Y_c(f) \sum_{-\infty}^{+\infty} \epsilon_T(f - nf_s)] \quad (28)$$

This relationship may be inverted to obtain an expression for the error in terms of the disturbance in the following way: Substitute for f ,

$$f^1 = f - m_{fs} \quad (29)$$

*The effect of the approximation disappears as $T \rightarrow \infty$, as we shall later do.

Then, for every integral m ,

$$\epsilon_T(f - m_{fs}) = Y_M(f - m_{fs}) \left\{ D_T(f - m_{fs}) - f_s Y_c(f - m_{fs}) \sum_{-\infty}^{+\infty} \epsilon_T[f - (n + m)f_s] \right\}. \quad (30)$$

But

$$\sum_{n=-\infty}^{+\infty} \epsilon_T[f - (n + m)f_s] = \sum_{n=-\infty}^{+\infty} \epsilon_T(f - n_{fs}). \quad (31)$$

Thus, if we write an infinite set of equations, using every integral value of m in the expression above once, and then sum,

$$\begin{aligned} \sum_{-\infty}^{+\infty} \epsilon_T(f - n_{fs}) &= \sum_{-\infty}^{+\infty} Y_M(f - n_{fs}) D_T(f - n_{fs}) \\ &- f_s \left[\sum_{-\infty}^{+\infty} Y_M(f - n_{fs}) Y_c(f - n_{fs}) \right] \sum_{-\infty}^{+\infty} \epsilon_T(f - n_{fs}). \end{aligned} \quad (32)$$

Hence,

$$\sum_{-\infty}^{+\infty} \epsilon_T(f - n_{fs}) = \frac{\sum_{-\infty}^{+\infty} Y_M(f - n_{fs}) D_T(f - n_{fs})}{1 + f_s \sum_{-\infty}^{+\infty} Y_M(f - n_{fs}) Y_c(f - n_{fs})} \quad (33)$$

and then from Eq. (28)

$$\begin{aligned} \epsilon_T(f) &= [1 - X(f)] Y_M(f) D_T(f) + [X(f)] \sum_{-\infty}^{+\infty} \\ &\times [Y_M(f - n_{fs}) D_T(f - n_{fs}) + Y_M(f + n_{fs}) D_T(f + n_{fs})] \end{aligned} \quad (34)$$

where

$$X(f) = \frac{f_s Y_c(f) Y_M(f)}{1 + f_s \sum_{-\infty}^{+\infty} Y_M(f - n_{fs}) Y_c(f - n_{fs})} \quad (35)$$

Using the methods outlined in Ref. 2, it follows from Eq. (34) that, if D is a stationary random time series having a spectral density function $\Phi_D(f)$, the spectral density of the error is given by:

$$\begin{aligned} \Phi_e(f) &= |1 - X(f)|^2 \left\{ |Y_M(f)|^2 \Phi_D(f) \right\} \\ &+ |X(f)|^2 \sum_{-\infty}^{+\infty} |Y_M(f - n_{fs})|^2 \Phi_D(f - n_{fs}) \\ &+ |Y_M(f + n_{fs})|^2 \Phi_D(f + n_{fs}) \end{aligned} \quad (36)$$

The integral of this function over all frequencies gives the mean-square error and may be so used to evaluate an existing or contemplated system. It is also obviously of the same form as the expression for error spectral density in the case of filtering independent signal and noise and, hence, Wiener's method for this case may be used directly to find the best X , given $Y_m(f)$, and $\Phi_D(f)$. In this case, Eq. (35) must be inverted to find $Y_c(f)$, the required optimum feedback frequency function. This can be done by the same technique used above to obtain Eq. (34) from Eq. (28). The result is:

$$Y_c(f) = \frac{1}{f_s Y_M(f)} \left\{ \frac{X(f)}{1 - \sum_{-\infty}^{+\infty} X(f - n_{fs})} \right\}.$$

REFERENCES

1. Wiener, N., "Extrapolation, Interpolation, and Smoothing of Stationary Time Series, with Engineering Applications," New York, John Wiley and Sons, 1949.
2. Stewart, R. M., "Statistical Design and Evaluation of Filters for the Restoration of Sampled Data," Proceedings of the I.R.E., 44 (No. 2), pp. 253-257, 1956.

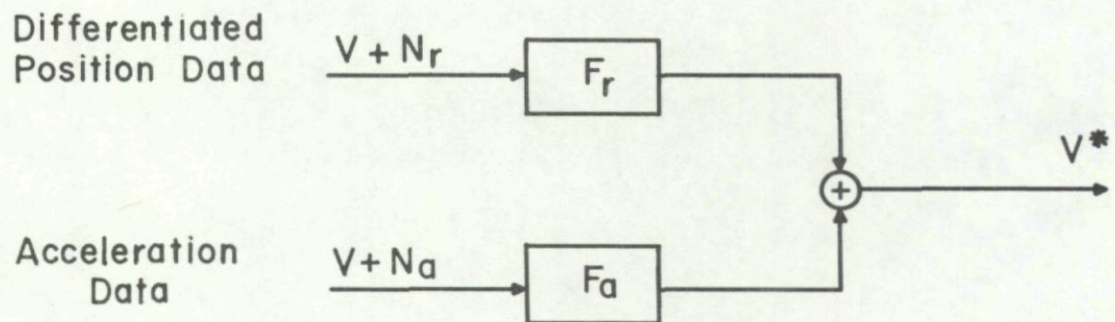


Fig. 1. Velocity from accelerometer and position data.

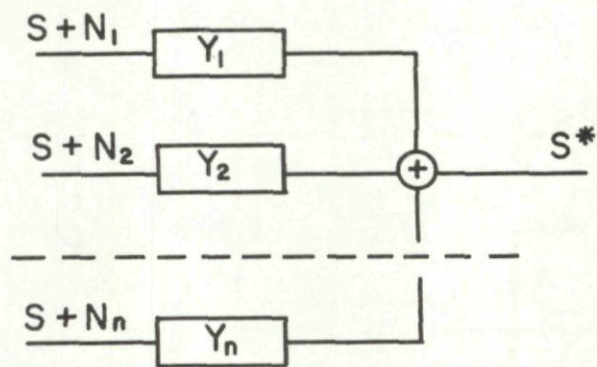


Fig. 2. Independent filtering of a series of measurements of a signal.

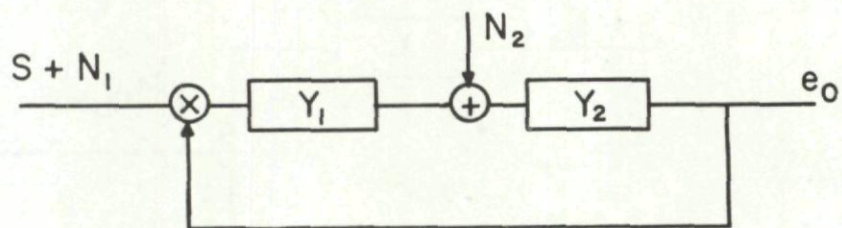


Fig. 3. Typical feedback system.

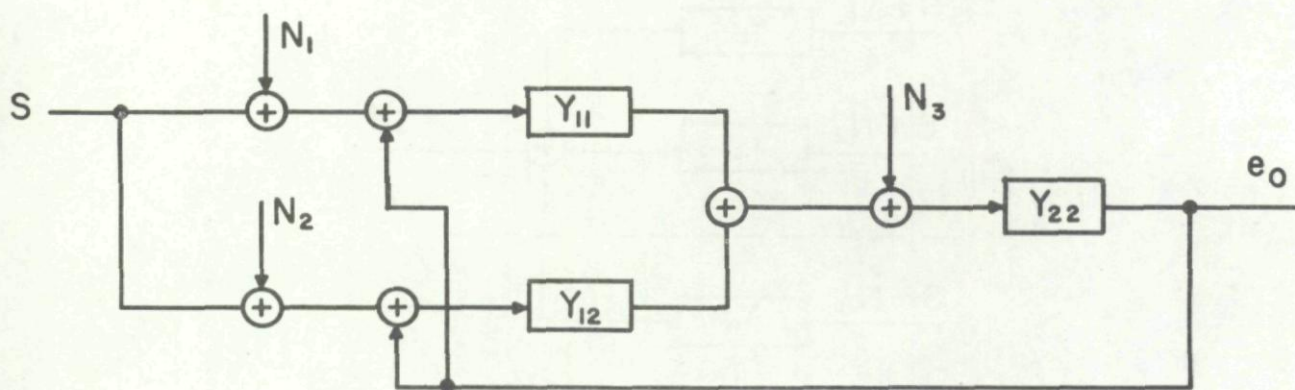


Fig. 4. Diagram of example.

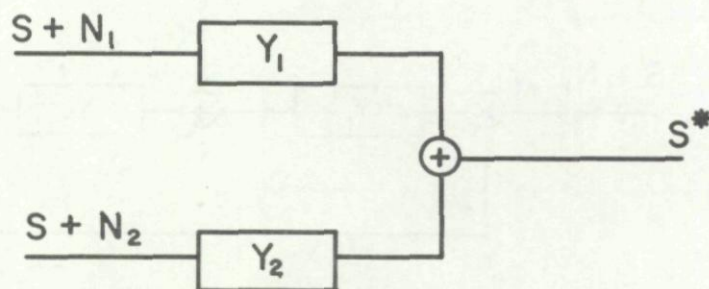


Fig. 5. Diagram of example.

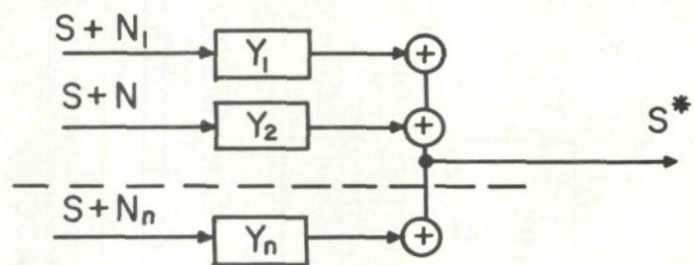


Fig. 6. Diversity system.

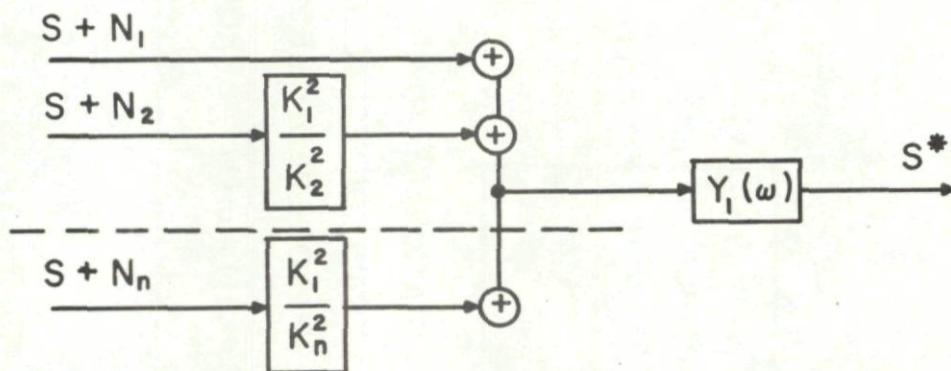
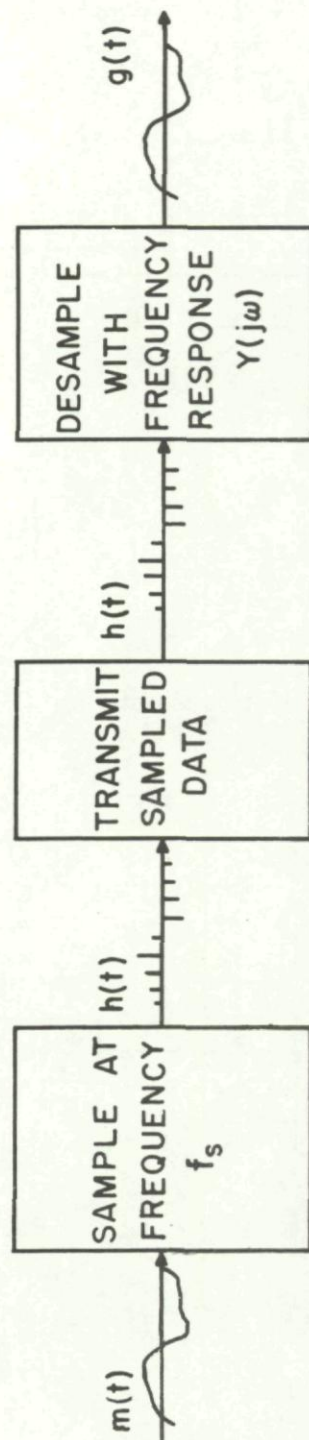


Fig. 7. Filtering solution of diversity problem.



$m(t)$ = STATIONARY RANDOM "MESSAGE" HAVING ZERO MEAN VALUE
 $h(t)$ = TRAIN OF EQUALLY SPACED SAMPLE PULSES OF FREQUENCY f_s
 $g(t)$ = RECOVERED MESSAGE
 $\epsilon(t) = g(t) - m(t)$ = ERROR IN RECOVERED MESSAGE

Fig. 8. Sequence of operations.

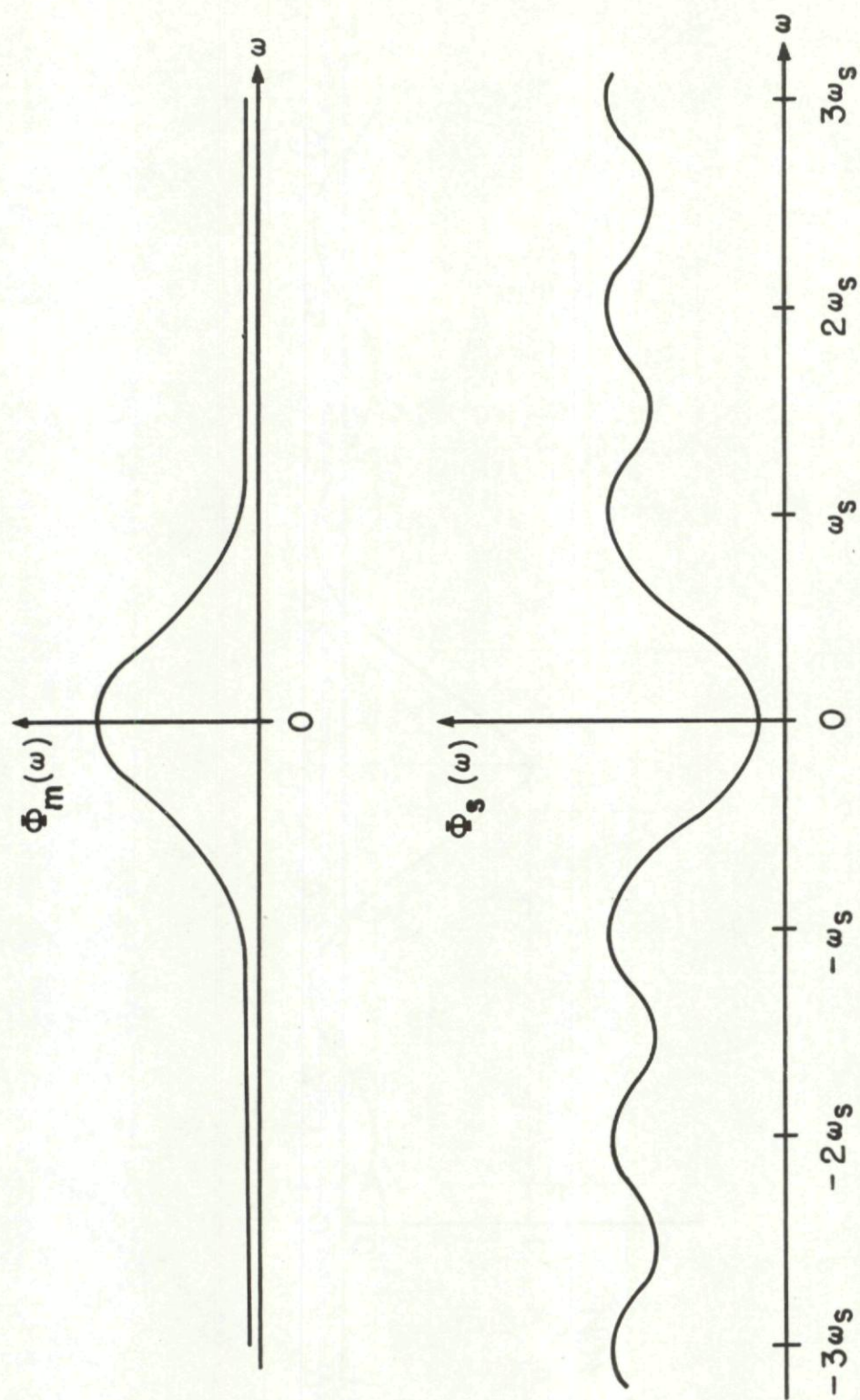


Fig. 9. Spectra of typical original message and of sidebands of effective noise.

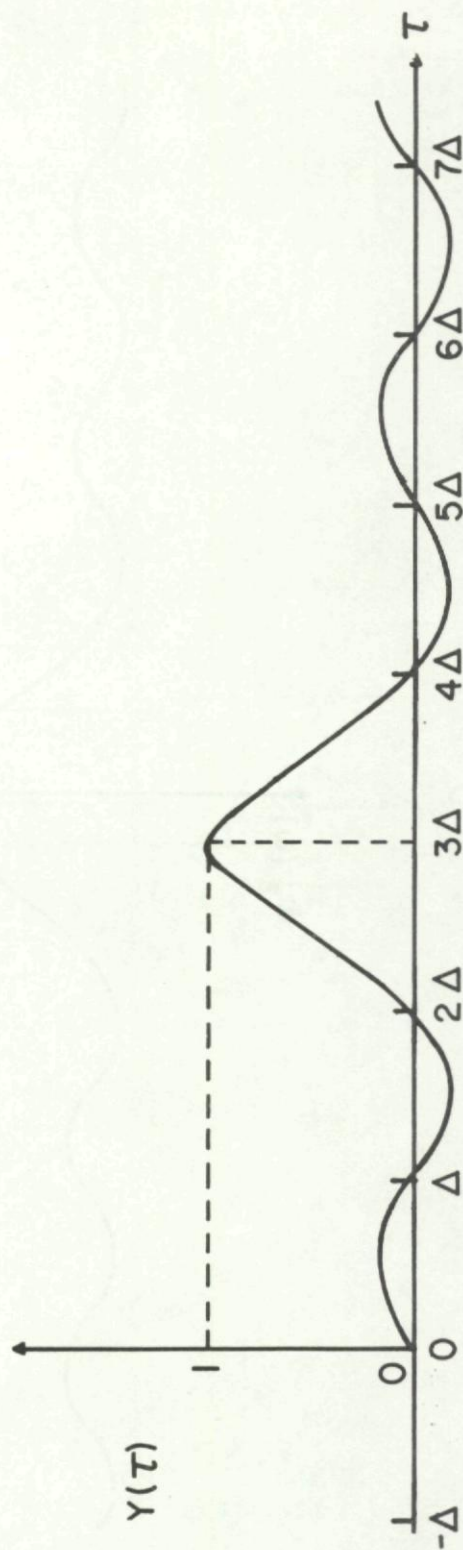


Fig. 10. Typical optimum weighting function for delay = 3Δ .

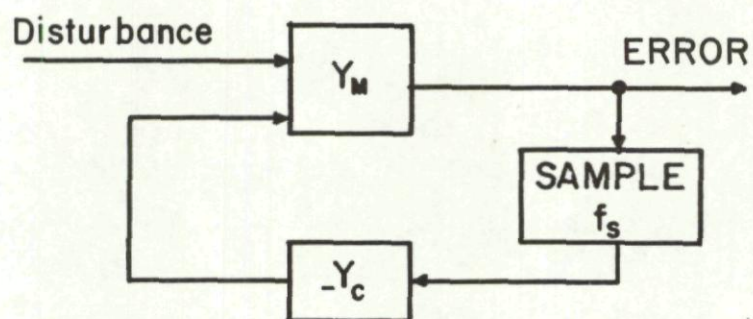


Fig. 11. Sampled error control system.



RECENT DEVELOPMENTS IN FIXED AND ADAPTIVE FILTERING

A. G. Carlton and J. W. Follin, Jr.*

SUMMARY

Classical optimal filtering methods have been extended to a large class of problems in which the input has incompletely specified characteristics. By minimax principles the optimal filter and the best input are determined. Two problems of time-varying filters are considered, first the optimal settling of filters to steady state and second the design of adaptive filters which adjust to varying or unknown environment.

SOMMAIRE

Les méthodes classiques de filtrage optimum ont été étendues à une large classe de problèmes, dans laquelle la grandeur d'entrée a des caractéristiques incomplètement spécifiées. En se basant sur des principes "minimax", le filtre optimum et la grandeur d'entrée la plus intéressante sont déterminés. Deux problèmes de filtres en fonction du temps, sont considérés, en premier l'ajustement optimum de filtre au régime permanent et deuxièmement l'étude de filtres adaptés qui ajustent au milieu inconnu ou variable.

1. INTRODUCTION

Linear filtering theory is largely based on the fundamental work of Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series," 1950, which in many respects paralleled the independent work of Kolmogorov.

It may be well to review the problem considered by Wiener. He assumed that there was available the entire past history of a time series consisting of signal-plus-noise possibly correlated with the signal, that all processes involved were stationary and indeed ergodic to the second order, with known autocorrelation and cross-correlation functions. He wished to determine the realizable linear filter to apply to the signal in order to minimize the mean-square difference between the output and the message translated by an assigned positive or negative time

interval. Wiener solved this problem by using variational analysis on the weighting function to obtain an integral equation, then by using subtle Fourier analysis to solve the integral equation. Multiple time series were handled by an extension of this technique.

It will be noted that the problem solved by Wiener contains two restrictions beyond the assumptions: first, optimization is restricted to linear functions; second, the loss function whose expectation is minimized is the squared error.

In attempting to extend filtering theory, it is appropriate to modify or eliminate various of these assumptions or restrictions. Various investigators have widened the field of permissible filters and dealt with alternative loss functions. Nonstationary processes have been considered; some trivial obvious results have been obtained, and some adaptive filters

*Applied Physics Laboratory, The Johns Hopkins University.

appear suitable, but little has been done that is both significant and rigorous. In this paper we shall remove the assumption of complete knowledge of the correlation functions and also indicate some minor extensions of the basic theory and techniques, consider the optimum filter with only portions of the signal history available, and attempt to classify the types of adaptive filters.

2. THE FREQUENCY APPROACH

The reader of Wiener's work will note that although his basic problem is formulated in terms of time series, correlation functions, and weighting functions, his solutions are expressed in terms of spectral densities and transfer functions. It appears reasonable, consequently, to set up the problem in the frequency domain.

From the elementary properties of spectral densities, we have

$$\sigma^2 = \int \left\{ |e^{i\omega\alpha} - F|^2 M + |F|^2 N \right\} d\omega. \quad (1)$$

where σ^2 is the mean-square error, $F(\omega)$ the filter transfer function, α the time translation, and M and N the signal-noise spectral densities, so normalized that the signal power is $\int M(\omega) d\omega$. All integrals are taken over the entire real frequency axis unless otherwise indicated, and the dependence of the variables on ω will usually not be indicated. It is assumed here and henceforth that the signal and noise are independent; this entails no loss in generality in the classical developments, where M can be regarded as the sum of the signal and signal-on-noise spectral densities, N as the sum of the noise and noise-on-signal spectral densities.

It is useful to consider the spectral densities as resolutions of the signal power into a continuum of frequency components. From this standpoint it is clear that ergodic properties are not relevant to the problem of linear filtering, although the optimal filter may be nonlinear if the second-order characteristics are not ergodic.

Before applying variations to minimize σ^2 by choice of F , let us indicate some extensions of this relation to problems beyond the original one. In the first place it will be noted that Eq. (1) is valid even though the power of signal, noise, or both is unbounded; it is not necessary that the correlation function exist. It will be seen that this may be of importance.

A trivial generalization of Eq. (1) is to replace $e^{i\omega\alpha}$ by the Fourier transform $Y(\omega)$ of any desired linear operator on the signal, e.g. by $i\omega$ for the derivative. Thus,

$$\sigma^2 = \int \left\{ |Y - F|^2 M + |F|^2 N \right\} d\omega. \quad (2)$$

Another easy extension is to the case in which it is desired to weight errors unequally for various frequency components. A symmetric nonnegative function $W(\omega)$, so normalized that $\int W(\omega) d\omega = 1$, could be inserted to obtain

$$\sigma_*^2 = \int \left\{ |Y - F|^2 M + |F|^2 N \right\} W d\omega, \quad (3)$$

which is of course formally equivalent to Eq. (2) with M and N replaced by MW and NW .

The next extension is to minimize σ^2 subject to a restriction on the mean power of the output or some other linear function of the

output. As an example, if σ^2 must be minimized subject to the restriction that the output acceleration power must be less than β , i.e., that

$$\int (M + N) |F|^2 \omega^4 d\omega \leq \beta, \quad (4)$$

we should minimize

$$\sigma^2 + \lambda \beta = \int \left\{ |Y - F|^2 M + |F|^2 N + \lambda (M + N) |F|^2 \omega^4 \right\} d\omega, \quad (5)$$

and select λ to satisfy Eq. (4). In certain cases, especially with non-Gaussian processes, restraints such as Eq. (4) may preferably be applied only to the noise. In this case the integral to be minimized can be reduced to Eq. (2) by suitable definition of N . This type of side condition can be introduced formally as here, or can be used in the definition of the class over which F is optimized. Several simultaneous side conditions can be introduced in the same way.

The final extension to be considered is to replace the class of realizable transfer functions by other classes, say \mathcal{F} , as appropriate.

3. MINIMIZATION OF σ^2 BY $F \in \mathcal{F}$

We now consider the selection of F within the class \mathcal{F} to minimize σ^2 . An increment J or F will produce an incremental σ^2 of

$$\begin{aligned} \sigma_J^2 + F - \sigma_F^2 &= \int |J|^2 (M + N) d\omega \\ &+ \int \bar{J} [(M + N) F - MY] d\omega, \end{aligned} \quad (6)$$

a relation obtained by noting that every transfer function has an even real part and an odd imaginary part.

The transfer function F will be optimal in \mathcal{F} if the right hand integral of Eq. (6) is nonnegative for every J such that $J + F \in \mathcal{F}$ and $\int |J|^2 (M + N) d\omega$ is finite.

The absolutely optimal F is, from Eq. (6), evidently

$$F_0 = \frac{MY}{M + N} \quad (7)$$

The optimal realizable F is given by

$$F_R = \frac{1}{(M + N)^+} \left[\frac{MY}{(M + N)^-} \right]_+ \quad (8)$$

where the new symbols denote factorization and decomposition of a meromorphic function as

$$H \approx H^+ H^- = H_+ + H_- , \quad (9)$$

H^+ being analytic and without poles or zeroes in the lower half plane, H^- the conjugate of H^+ ; H_+ and H_- have no poles in the lower half plane and upper half plane, respectively. Polynomial terms of H appear in H_+ . If $M + N$ is not factorable, $F_R = F_0$.

To check the validity of Eq. (8), substitute it in Eq. (6), obtaining

$$\begin{aligned} \sigma_J^2 + F_R - \sigma_{F_R}^2 &= \int |J|^2 (M + N) d\omega \\ &+ \int \bar{J} (M + N)^- \left[\frac{MY}{(M + N)^-} \right]_- d\omega; \end{aligned} \quad (10)$$

the latter integral is zero, by contour integration over the upper half plane, for J sufficiently convergent, i.e., for $\int |J|^2 (M + N) d\omega$ finite.

As an example let us consider the spectra

$$M = \theta/\omega^4, \quad N = \phi, \quad \text{with } Y = 1, \quad (11)$$

the absolute optimum (realizable with infinite delay) is then

$$F_0(p) = \frac{1}{1 + \phi p^4/\theta} \quad (12)$$

and the optimal realizable F is

$$F_R(p) = \frac{1 + \sqrt{2} p (\phi/\theta)^{1/4}}{1 + \sqrt{2} p (\phi/\theta)^{1/4} + p^2 (\phi/\theta)^{1/2}} \quad (13)$$

This transfer function is the zero-velocity lag loop with .7 critical damping which will be discussed below. The corresponding errors are

$$\begin{aligned} \sigma_0^2 &= .3535 \phi^{3/4} \theta^{1/4} \\ \sigma_R^2 &= 1.414 \phi^{3/4} \theta^{1/4}. \end{aligned} \quad (14)$$

A very useful relation can be obtained by manipulation of the second integral of Eq. (6) as follows:

$$\begin{aligned} &\int \bar{J} [(M+N)F - MY] d\omega \\ &= \int \bar{J}/\bar{F} [(M+N)|F|^2 - MY\bar{F}] d\omega \\ &= \int \bar{J}/\bar{F} [(M+N)|F|^2 - \text{Re} MY\bar{F} - \text{Im} MY\bar{F}] d\omega. \end{aligned} \quad (15)$$

This integral is zero for sufficiently convergent realizable J if

$$|F_R|^2 (M+N) = \text{Re} MY\bar{F} + \text{Re:Im} MY\bar{F}, \quad (16)$$

where $\text{Re:Im}H$ represents the real complement to the given imaginary function such as to render H realizable.

This can be more formally expressed by the Bode relation

$$|F_R|^2 (M+N) = \frac{1}{i\pi} \int_* \frac{w(MY\bar{F})_w dw}{\omega^2 - w^2}, \quad (17)$$

where \int_* indicates disregard of poles at $w = \pm \omega$, provided MY does not diverge for large ω .

Solutions of Eq. (16) for simple forms of MY are readily obtained; for example,

$$|F_R|^2 (M+N) = \begin{cases} MY, & \text{if } MY \text{ is constant} \\ MY\bar{F}(ia), & \text{if } MY = \frac{c \cdot a^2}{a^2 + \omega^2} \\ MY + k, & \text{if } MY = c_0 + c_2 \omega^2 + c_4 \omega^4 \end{cases} \quad (18)$$

$$[K \text{ determined by } \int \log |F_R|^2 d\omega = 0].$$

By symmetry under the interchange of $M \leftrightarrow N$ and $F \leftrightarrow (Y - F)$ we obtain

$$|Y - F_R|^2 (M+N) = NY, \quad (19)$$

if NY is constant.

For ordinary filtering, with $Y = 1$, these define the optimal spectral density of the error signal in a servo type filter, and can be used to construct an adaptive filter which will become optimal for any signal spectrum.

Expressions for $|F_R|^2$, based on Eq. (16) are quite useful in optimizing filters with side conditions such as limited mean-square output, since the expressions for $|F_R|^2$ are frequently more convenient to use than those for F_R in determining the Lagrange multipliers.

4. MINIMAX FILTERING

Let us consider now some typically statistical problems, in which one has incomplete knowledge of the spectrum of noise, of signal, or both. We discuss below adaptive quasi-linear filters which appear suitable for cases in which one spectrum is completely unknown, and which can cope with cases involving a spectral density of known form but unknown magnitude.

The problems considered by Wiener were essentially probabilistic, i.e., the system is completely described in terms of appropriate probability measures; the problems we are now considering are statistical, in that we are dealing with a system defined by probabilities, some of which are unknown. Our problem in this case is one of statistical estimation of a function of the signal. Our estimating function should be optimal in some sense. One of the most logical criteria for an estimate, developed by Abraham Wald, is that it minimizes the maximum expected loss; that is, each filter is assessed on the basis of the expected loss with the possible system which is least favorable for the given filter, and the optimal filter is that filter for which the maximum expected loss is minimum. This formulation of statistical decision theory is very similar to two-person game theory, independently developed by von Neuman. We adopt this criterion and consider as optimal the minimax filter, with the loss function proportional to the squared error. In a completely prescribed system, the minimax linear filter is the Wiener optimal linear filter.

Minimax theory offers a strong justification for the use of linear filters. If the distribution functions of the processes are not known but the class of possible distributions includes Gaussian distributions, the minimax filter is linear, since with a

linear filter the mean-square error is independent of the form of the distribution function, and with a nonlinear filter the mean-square error exceeds that with a linear filter when the processes are Gaussian.

Typical problems encountered in practice involve situations in which the noise process is known to be limited in power, in mean square velocity, etc.:

$$\int N d\omega = C_0, \int N\omega^2 d\omega = C_1, \text{ etc.} \quad (20)$$

Such restrictions can be put in the general form

$$\int N \theta d\omega = 1, \quad (21)$$

where θ is a prescribed symmetric non-negative function. It can be shown straightforwardly that the variation in σ^2 due to variation n in N is a nonnegative function of n , zero if and only if $n = 0$, plus

$$\int n \left[|F_{M,N}|^2 - \lambda \theta \right] d\omega, \quad (22)$$

where $F_{M,N}$ is the optimal transfer function with M and N , and λ is a Lagrange multiplier to be selected to satisfy Eq. (21). From Eq. (22) and the fact that $n + N$ must be non-negative, it follows that the maximum N is N_0 given by

$$\begin{aligned} |F_{M,N_0}|^2 &= \lambda \theta & \omega: N_0 > 0 \\ &\leq \lambda \theta & \omega: N_0 = 0. \end{aligned} \quad (23)$$

This result is easily generalized to the case of several inequality restrictions

$$K_j \int N \theta_j d\omega = 1 \quad (j = 1, 2, \dots, k) \quad (24)$$

with the K_j not specified but ≥ 1 .

The maximum N is N_0 satisfying

$$\begin{aligned} |F_{M,N_0}|^2 &= \sum \lambda_j K_j \theta_j & \omega: N_0 > 0 \\ &\leq \sum \lambda_j K_j \theta_j & \omega: N_0 = 0. \end{aligned} \quad (25)$$

with the λ_j , K_j satisfying Eq. (24) and also

$$\lambda_j (K_j - 1) = 0, \quad (j = 1, 2, \dots, k). \quad (26)$$

Similar results are obtained for cases in which the signal spectral density is subject to one or more equalities or inequalities, and where both spectra are limited only by such restrictions.

The results just given have derived the maximin spectrum or spectra, but our object is to determine the minimax filter. For this purpose we now prove that

$$\min_F \max_N \sigma_{N,F}^2 = \max_N \min_F \sigma_{N,F}^2$$

and that the minimax F is F_{M,N_0} .

To prove this, we observe that

$$\begin{aligned} \min_F \max_N \sigma_{N,F}^2 &\leq \max_N \sigma_{N,F_{M,N_0}}^2 \\ &= \max_N \int \left\{ N |F_{M,N_0}|^2 + M |\gamma - F_{M,N_0}|^2 \right\} d\omega \\ &= \sum \lambda_j + \int M |\gamma - F_{M,N_0}|^2 d\omega = \max_N \min_F \sigma_{N,F}^2. \end{aligned}$$

We have thus shown that

$$\min_F \max_N \sigma_{N,F}^2 \leq \max_N \min_F \sigma_{N,F}^2. \quad (27)$$

But by the fundamental theorem of game theory,

$$\min_F \max_N \sigma_{N,F}^2 \geq \max_N \min_F \sigma_{N,F}^2. \quad (28)$$

from which it follows at once that the "game" is determined, with the minimax filter being F_{M,N_0} and the maximin N being the N_0 previously defined.

The basic result of the minimax approach to optimum filtering is that the errors depend in the second order on the spectra and the form of the filter. As a consequence, if a suitable approximation to the optimal form is used and the parameters are adjusted properly the resulting system will be satisfactory.

5. TIME-VARYING FILTERS

Let us now consider the problem of filtering when only a finite and perhaps fragmentary history of the signals is available. In this case the filter parameters are variable, and we must assume a particular form of transfer function. In general, the steady-state optimal filter with variable band-pass and damping is best. We may attack this problem in the time domain by considering the rate of change of the filtering errors and adjusting parameters to maximize the rate of decrease of the error.

For a linear system the tracking accuracy may be described in terms of the variances of the tracking error and error rate. As an example let us consider the simple zero

velocity lag feedback system in Fig. 1. The input consisting of signal and noise $x(t) + x_n(t)$ is at the left, the output, x_c , at the right. The equations of the system are

$$\begin{aligned}\frac{dx_c}{dt} &= \dot{x}_c + b(x - x_c) + bx_n \\ \frac{d\dot{x}_c}{dt} &= a(x - x_c) + ax_n\end{aligned}\quad (29)$$

Note, in explanation of the notation, that $\dot{x}_c \neq \dot{x}_c$.

Actually the principal interest centers upon the errors $\epsilon_c = x - x_c$, $\epsilon_{\dot{c}} = \dot{x} - \dot{x}_c$. In the second diagram of Fig. 1 is shown the error loop equivalent to the original signal loop. The signal x now appears as an acceleration input to the first integrator. This is a significant advantage when, as is often the case, the acceleration spectrum of the signal is known.

So far a , b are unrestricted. If \ddot{x} and x_n both have flat spectral densities θ , ϕ respectively then this filter is optimal with the values of a and b previously derived. The present purpose is to extend the optimization to the transient period. The gains a , b in this case are time functions and the resulting system, while not necessarily optimum among all possible systems, is the best obtainable with a given structure.

The key to the solution lies in setting up the differential equations relating the variances and covariances of the integrator outputs ϵ_c , $\epsilon_{\dot{c}}$. Write Eqs. (29) in terms of the errors and express the solution in the neighborhood of t as a power series in Δt . Terms beyond the first degree in Δt are not

required. The result is most easily obtained by direct use of the second figure

$$\epsilon_c = \epsilon_{c_0} + \epsilon_{\dot{c}_0} \Delta t - b \epsilon_{c_0} \Delta t + b \int_t^{t+\Delta t} x_n dt + O(\Delta t^2) \quad (30)$$

$$\epsilon_{\dot{c}} = \epsilon_{\dot{c}_0} - a \epsilon_{c_0} \Delta t + \int_t^{t+\Delta t} \ddot{x} dt + a \int_t^{t+\Delta t} x_n dt + O(\Delta t^2). \quad (31)$$

Square Eq. (30) and average over the ensemble of inputs \ddot{x} , x_n . Then denoting resulting variances and covariances by $\widehat{\epsilon_c}$, $\widehat{\epsilon_{\dot{c}}}$, $\widehat{\epsilon_c \dot{c}}$, respectively

$$\widehat{\epsilon_c} - \widehat{\epsilon_{c_0}} = 2 \widehat{\epsilon_{c_0} \dot{c}} \Delta t - 2b \widehat{\epsilon_{c_0}} \Delta t + b^2 \phi \Delta t + O(\Delta t^2) \quad (32)$$

where ϕ is the spectral density (assumed constant of x_n).

The term ϕ is derived as follows: $\phi(a)$ is the autocorrelation function of x_n .

$$\begin{aligned}& \left[\int_t^{t+\Delta t} x_n dt \int_t^{t+\Delta t} x_n dt \right] \\ &= \int_t^{t+\Delta t} \int_t^{t+\Delta t} [x_n(u) x_n(v)] du dv \\ &= \int_0^{\Delta t} \int_0^{\Delta t} \phi(u-v) du dv \\ &= \phi \int_0^{\Delta t} \int_0^{\Delta t} \delta(u-v) du dv \\ &= \phi \Delta t.\end{aligned}$$

Dividing Eq. (32) by Δt and letting $\Delta t \rightarrow 0$

$$\frac{d\widehat{\epsilon_c}}{dt} = 2 \widehat{\epsilon_{c\dot{c}}} - 2b \widehat{\epsilon_c} + b^2 \phi. \quad (33)$$

Similarly, assuming the spectral density θ of \ddot{x} is flat

$$\frac{d\hat{\epsilon}_{\dot{c}}}{dt} = -2a \hat{\epsilon}_{\dot{c}\dot{c}} + a^2 \phi - \theta \quad (34)$$

$$\frac{d\hat{\epsilon}_{\dot{c}\dot{c}}}{dt} = \hat{\epsilon}_{\dot{c}} - a \hat{\epsilon}_{\dot{c}} - b \hat{\epsilon}_{\dot{c}\dot{c}} + a b \phi. \quad (35)$$

These variance and covariance equations may be used to adjust the gains to get optimum tracking. This will surely result if functions a , b can be found making the right hand members of Eqs. (33), (34), and (35) simultaneously minimum for this will make $\hat{\epsilon}_{\dot{c}}$ and $\hat{\epsilon}_{\dot{c}\dot{c}}$ decrease at maximum rate.

This simultaneous minimum does occur and at

$$a = \frac{\hat{\epsilon}_{\dot{c}\dot{c}}}{\phi}, \quad b = \frac{\hat{\epsilon}_{\dot{c}}}{\phi}, \quad (36)$$

as can be seen by setting the partial derivatives of all three right members with respect to a and b equal to zero. The resulting system has optimum tracking and rate of settling and the variances facilitate evaluating performance of the system.

This optimization of the transient behavior has a byproduct, the known steady-state result. For in this case the left hand members of Eqs. (33), (34), and (35) vanish when using Eq. (36)

$$a = \left(\frac{\theta}{\phi}\right)^{1/2}, \quad b = \sqrt{2} \left(\frac{\theta}{\phi}\right)^{1/4}. \quad (\text{steady-state}) \quad (37)$$

For simplicity the input acceleration spectrum has been assumed flat. There are several ways to deal with nonflat spectra. For example it can be shown that for a general $\theta(\omega)$ with autocorrelation function $\phi(t)$, θ in Eq. (34) would be replaced by $2 \int_0^t W(t, \tau) \phi(t - \tau) d\tau$. Here $W(t, \tau)$ is the impulsive response of $\epsilon_{\dot{c}}$ to \ddot{x} . Or a flat spectrum could be filtered by $\theta(\omega)^* [\theta(\omega)^- = \theta(\omega)^* \theta(\omega)^-]$ to give a signal of spectrum $\theta(\omega)$ into the integrator.

This last is particularly simple in the Markoff case $\theta = \theta_0 / (C^2 + \omega^2)$ where $\theta^* = 1 / (C + j\omega)$. The original error tracking loop could be modified as shown in the third figure thus introducing one additional integrator. Proceeding as before, six variance-covariance equations result. In general, with a system involving n integrators (n 'th order differential operation), there will be one-half $n(n + 1)$ variance equations although generally some are trivial.

The transient filtering problem has been discussed on the basis that the noise and signal spectra are known. This leads to a solution of the optimum settling time of a filter and to the best combination of parameters even if the transfer function is not optimal. If the noise and signal spectra are not known then the above technique of computing variances fails and other methods must be used. If sufficient time is available it is possible to measure the spectral densities and use the variance methods, but less cumbersome methods are desirable.

6. ADAPTIVE FILTERS

The discussion which follows will concern adaptive systems, i.e., those which change parameters or adapt as a function of the environment. In general, the rate of change of parameters is slow compared to the data rate of the input so that they may be treated

as time-varying rather than nonlinear systems. If the response of the adapting loop is fast, paper analysis is impossible and simulation techniques are needed. In designing an adaptive system it is necessary to consider the response and stability of the loop as well as the source of intelligence to be employed in adjusting the system.

Adaptive systems may be classified according to several distinct criteria as follows:

a. Object of adaptive loop

- (1) Setting of gain or transfer function.
- (2) Adjusting output level or other parameter.
- (3) Adjusting stability margins of main loop.

b. Source of information

- (1) Measurements of normal input or output.
- (2) Injection of tracer signal outside band-pass of normal input.
- (3) Time sharing tracer signal.
- (4) Amplitude or phase of self-excited oscillations.

c. Type of system

- (1) Open loop, i.e., system adjusted according to measurements on the input.
- (2) Closed loop, i.e., system measures signal or tracer output.

In addition, all adaptive systems may be classified according to standard servo practice as electrical or mechanical, digital or analog, etc., but such distinctions are not desirable for the present purpose. The class of adaptive servos ranges from standard AGC and AFC loops to servo driven auto-transformers for voltage regulation to more sophisticated optimal filtering loops.

Let us now look at examples of the three different adaptive systems listed under the first criterion (a). These are the zero velocity lag tracking loop mentioned earlier, a similar filter with limited output acceleration, and a system for maintaining a servo loop as tight as possible without instability when the loop gain is slowly varying or not known accurately.

Fig. 2 shows a simple servo where only the gain is varied and the loop gain is to be maximized (possibly to minimize the effects of a variable back torque from the load). The unknown gain is assumed to be in the servo. The band-pass filter passes the frequency at which instability is expected and this signal is detected and used to adjust the gain to damp the oscillations.

Let λ_0 be the gain at which the loop is neutrally stable. Then the rate of buildup or decay of oscillations is proportioned to $(\lambda/\lambda_0 - 1)$. The amplitude, z , of the oscillations satisfies the equation

$$\dot{z} = k_1 z(\lambda - \lambda_0)/\lambda_0 \doteq k_1 z_0 (\lambda - \lambda_0)/\lambda_0. \quad (38)$$

If the signal is picked off at point A, the control equation is

$$\dot{\lambda} = -g(s) (z - z_0), \quad (39)$$

while if it is picked off at point B, we have

$$\begin{aligned}\dot{\lambda} &= g(s) \lambda (z - z_0) \\ \dot{\lambda} &= -\lambda_0 g(s) (z - z_0).\end{aligned}\quad (40)$$

Combining these equations we find that in case A

$$\left[s^2 + \frac{k_1 z_1}{\lambda_0} g(s) \right] \lambda = -k_1 z_1 \lambda_0 g(s) \left(\frac{1}{\lambda_0} \right), \quad (41)$$

while for case B

$$\left[s^2 + k_1 z_1 g(s) \right] \lambda = k_1 z_1 g(s) \lambda_0. \quad (42)$$

While both systems can be made stable, and both have $\lambda = \lambda_0$ as the steady-state solution, the transient response of the adaptive loop depends on the required gain in case A, but in case B the transient is fixed and the system has zero velocity lag with respect to variations of λ_0 . If λ_0 increases linearly with time then $\lambda = \lambda_0$ but $z > z_1$ with $\dot{z} = 0$. Thus a closed-loop adaptive system is better here. These statements are subject to modification if there exists dead space or friction in the main or adaptive loop; actually simulation is then required to determine the behavior.

If the main loop shaping network $f(s)$ is properly chosen then a very good servo response is obtainable but input signals at the loop resonance frequency must be avoided. This design is especially useful for a regulator, i.e., when $x_i = 0$ and the servo is designed to counter the back torque.

Fig. 3 is a diagram of an adaptive filter in which the r.m.s. output acceleration due to noise is limited. The adaptive loop is very simple and, with the gains as shown, has a response which is independent of input noise spectral density. This can be

seen from the fact that the error in r.m.s. acceleration is proportional to $\Delta \lambda / \lambda$ and hence the control equation is

$$\dot{\lambda} \sim \Delta \lambda, \quad (43)$$

if the filter $g(s)$ is unity.

If it is desired to have the adaptive loop time constant proportional to the main loop time constant then the λ must be replaced by λ^2 and the control equation is

$$\dot{\lambda} \sim \lambda \Delta \lambda. \quad (44)$$

If the noise is flat or of known spectral shape it is possible to measure its amplitude at the input to the filter and compute the r.m.s. acceleration from the known transfer function of the filter. It is then possible to use an open loop method of adjusting λ but, while this eliminates adaptive loop stability problems it does not have the accuracy in adjusting λ . If the noise does not have the expected spectrum the filter $f(s)$ will not be optimal in shape; however the performance of the system will deviate in the second order if λ is correct, but errors in λ may give first order effects on system performance due to violation of the constraints.

While this loop is very simple we have discussed it because some simulator studies of the effects of nonlinearities may be of interest. The nonlinearity concerned is that due to a fixed limiter inserted in the loop as shown at the bottom of Fig. 3. The analysis was carried out using various fixed values of λ and noise rather than closing the adaptive loop as above.

Fig. 4 shows the results obtained. The solid upper curve shows σ^2 vs. λ for an unlimited system. If the limiter is inserted then the dashed curve is appropriate; the minimum is only a few percent above the optimum at the minimum. The lower half of the figure shows the effect of the limiter

on the output acceleration before and after the limit. The result of the simulator study was that the minimum in σ^2 occurs almost exactly at $a_1 = L$, hence the simple adaptive system just described forms a very sophisticated tracking loop.

Fig. 5 is a block diagram of the zero velocity lag tracking loop in which we do not know the signal or noise spectra although we assume the noise to be nearly flat. In order to adjust band-pass we may use the result, Eq. (18) that the error signal spectrum is proportional to the noise spectrum when the loop is optimal. While the transfer of the loop is not correct if the noise is not flat or if the signal is not that assumed, it is still true that adjusting the loop band-pass so that the error signal spectrum is flat is nearly optimum.

The method of measurement is to use two filters $f_1(s)$, a low-pass filter covering the band-pass of the main loop, and $f_2(s)$ covering an equal band-pass just above the main loop and take the ratio of the outputs. The optimum shape of such filters has not been determined but simulator runs show suitable performance for simple filters. $f_2(s)$ should have a finite band-pass because the actual high frequency noise is unimportant; only the noise in the vicinity of the main loop band-pass is important.

In the loop as shown the obvious scale factors have been inserted to make the response frequency (relative to the main loop) independent of the value of the spectra. The filter $f_3(s)$ determines the band-pass of the adaptive loop and, in order to have minimum r.m.s. errors in λ , $f_3(s)$ must be adjusted so that the lags in following changes in the spectra are balanced by the fluctuations in the noise out of the detectors. From this criterion we can determine the adaptive loop band-pass as $\omega_A \sim \sqrt{\lambda/T}$, where $1/T \sim \ddot{\phi}/\dot{\phi}$ is the effective time constant relating to the change in input spectra. However, if step

changes in the input signals are contemplated then the adaptive loop should be as tight as stability dictates and the gain settings in the figure are correct.

The ratio of the filter outputs minus one is proportional to $\delta \lambda / \lambda$ so that

$$\dot{\lambda} \sim \lambda \Delta \lambda \quad (45)$$

and the band-pass in the adaptive loop is proportional to that in the main loop.

In all of the adaptive systems considered it is easy to specify the gain changes to keep the loop dynamically similar for different inputs, but it is harder to specify the exact band-pass or the shape of the filters in the adaptive loop. It is always possible to make a linear stability analysis, if the inputs are fixed, and the noise out of the squaring circuits can be computed for Markovian noise but no general theory of optimal design exists.

There are many ways of instrumenting adaptive servos which give adequate performance and the effects of nonlinearities and complexity must be considered carefully if a satisfactory design is to be obtained. For example, the use of smoothed absolute value instead of r.m.s. leads to only a few percent more noise in the adaptive loop. As another example the division in the last example may be replaced by a subtraction if the dynamic range is not too great. At low input signals the loop is then sluggish but the tracking error is small due to the small input.

Underlying the design of adaptive servos is the assumption of relatively slow, or only occasional, changes in environment. If rapid changes occur a different main loop, possibly nonlinear, is required. Adaptive loops may be called quasi-linear but because they are nonlinear no general method of analysis has emerged to determine optimal performance as a standard of comparison with specific loops, or to check instrumentation approximations.

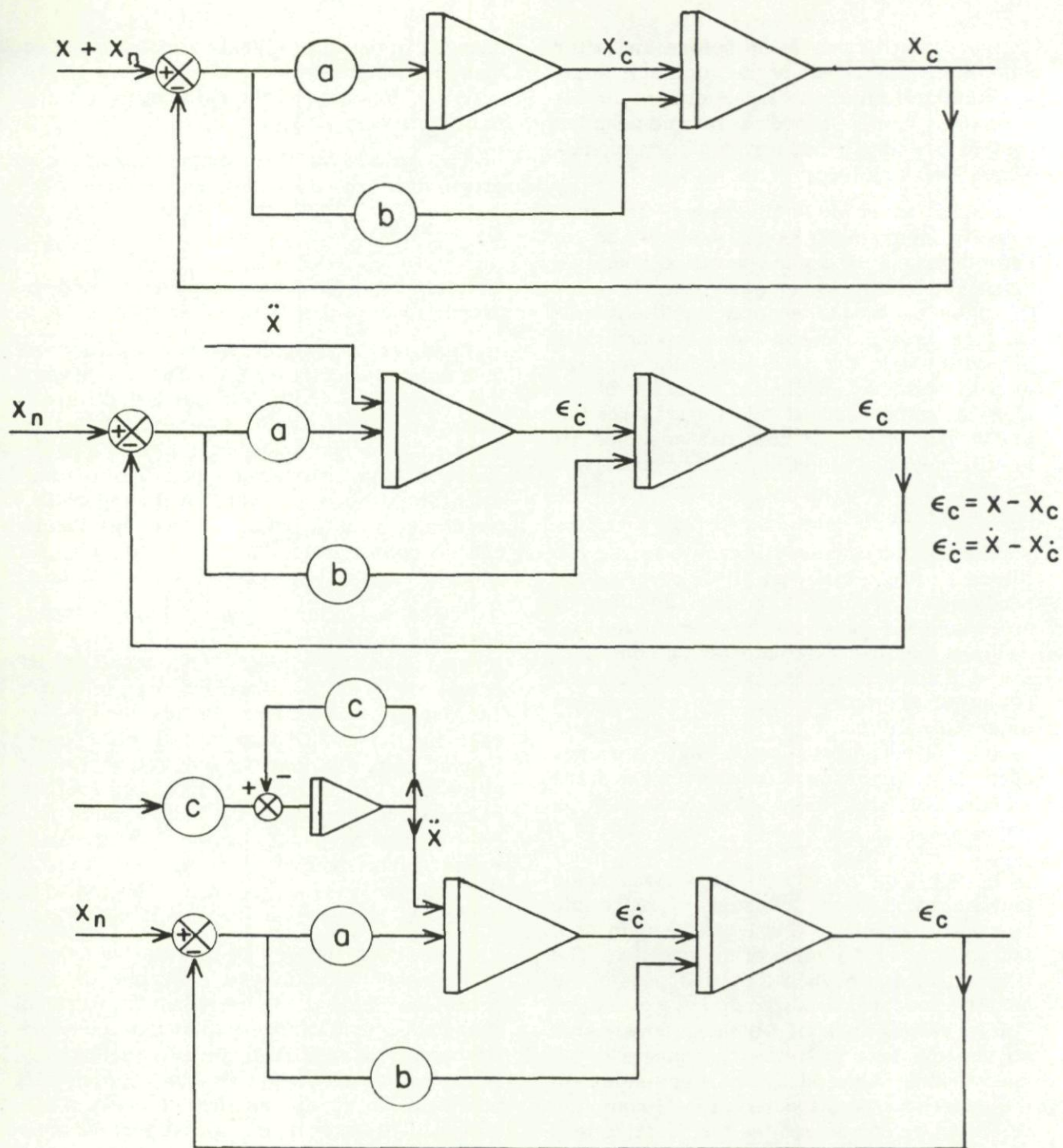


Fig. 1. Zero velocity lag feedback system.

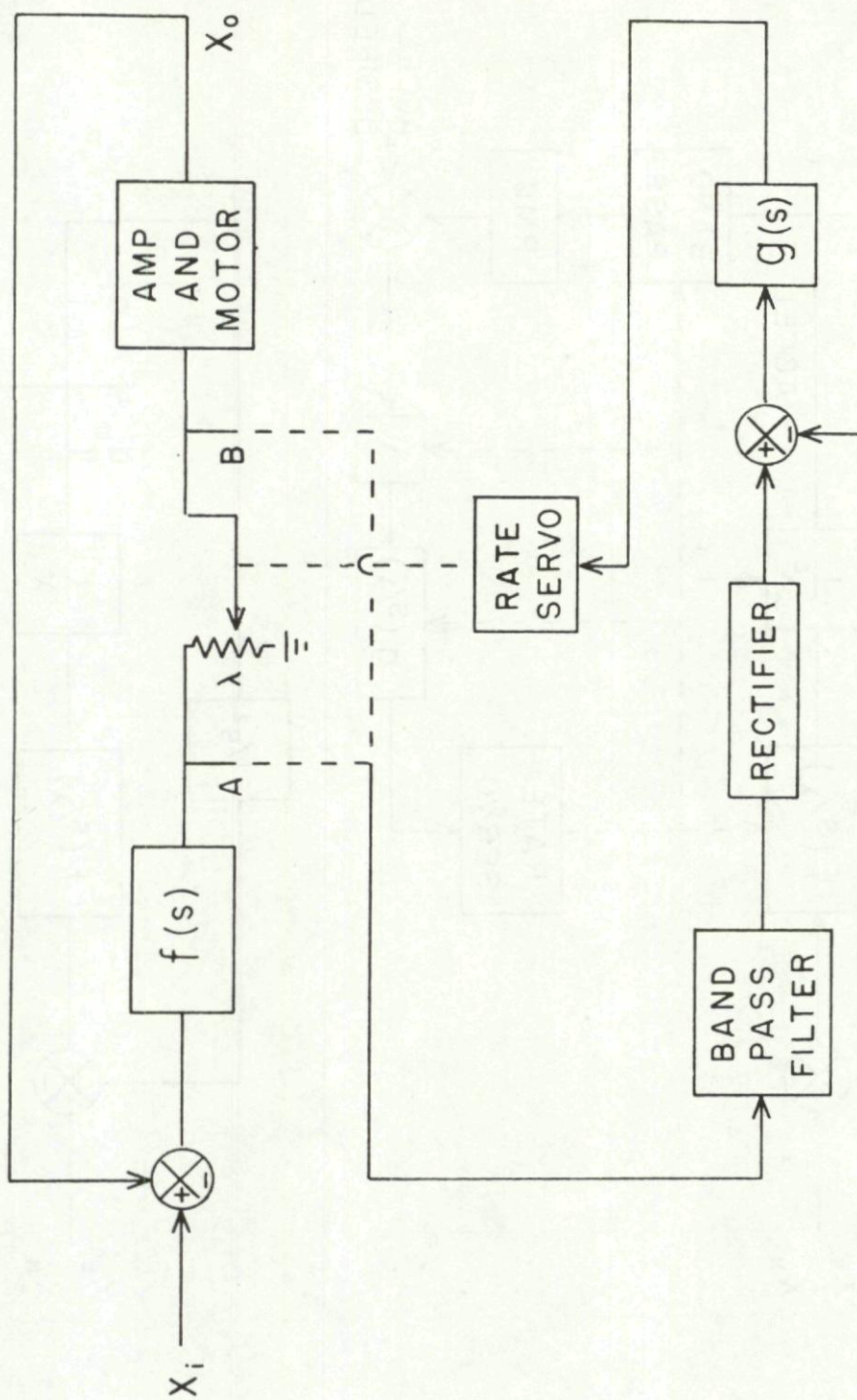


Fig. 2. Maximized loop gain servo.

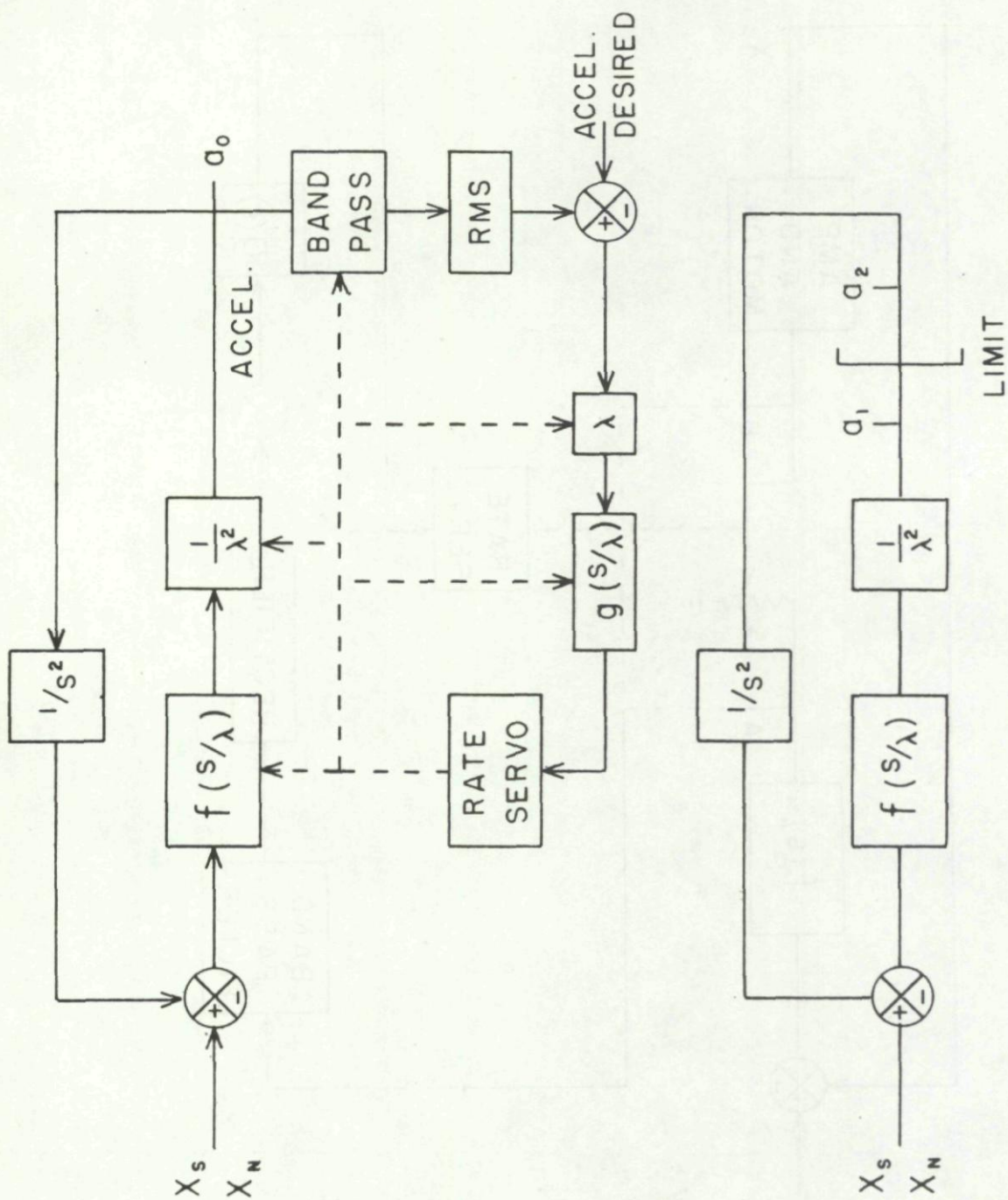


Fig. 3. An adaptive filter.

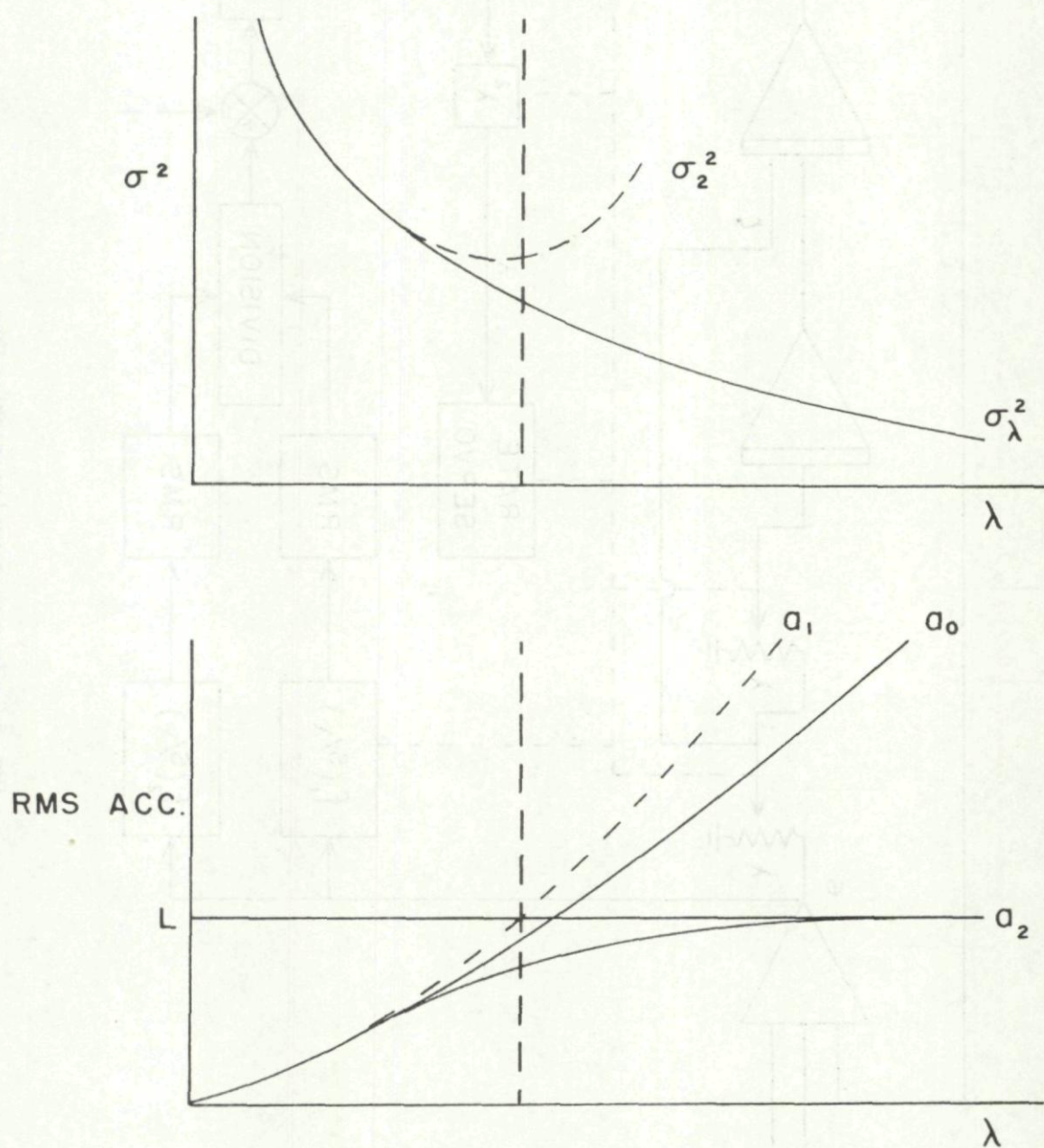


Fig. 4. Results of system of Fig. 3

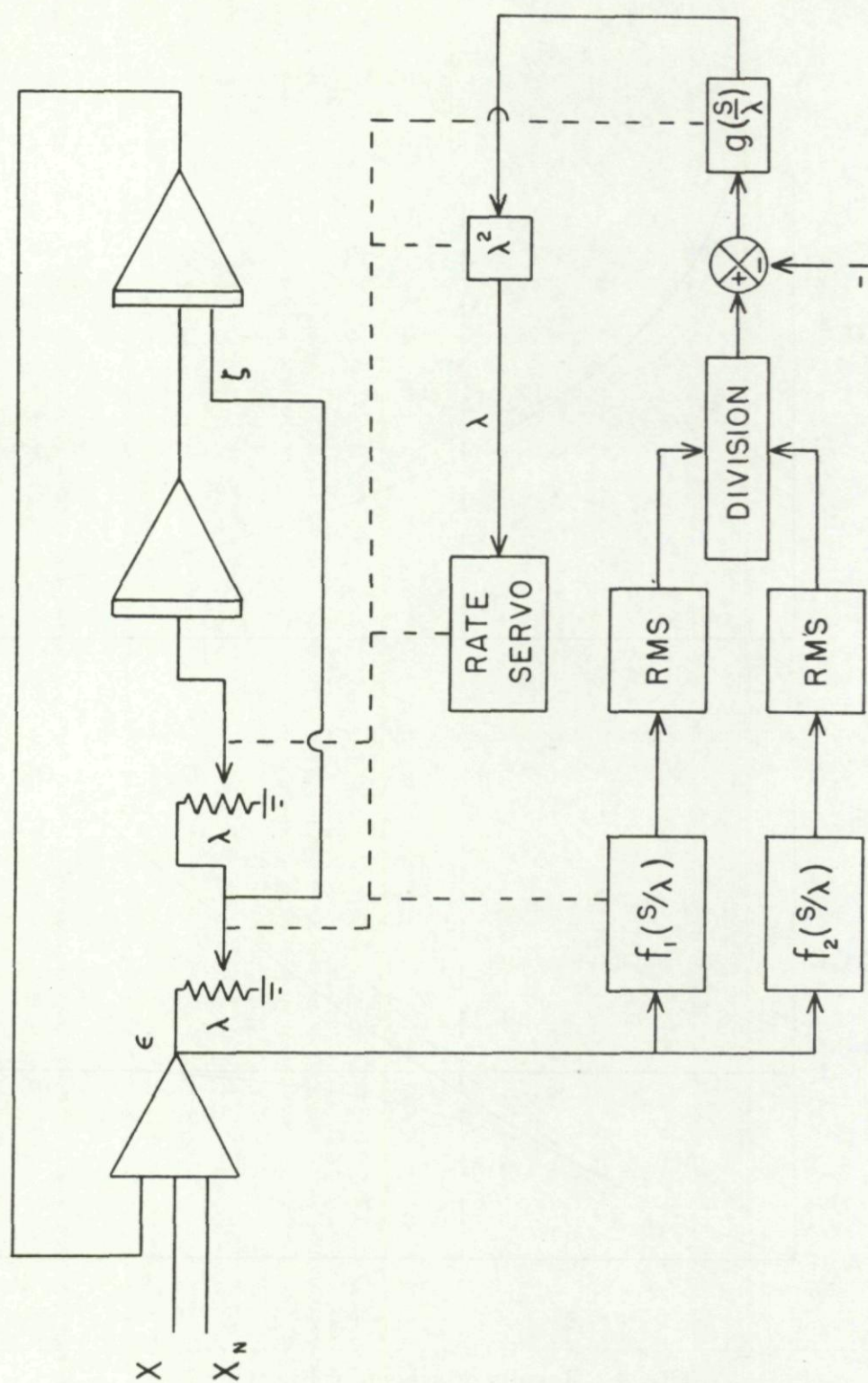


Fig. 5. Zero velocity lag tracking loop.

PRACTICAL PROBLEMS ENCOUNTERED IN MISSILE GUIDANCE AND CONTROL DESIGN

R. E. Whiffen*

SUMMARY

This paper discusses various problems to be overcome during the process of designing, making, ground testing, and successfully flying missiles. The importance of reliability is stressed. Emphasis is on designing and testing the missile itself, specifically the electronic parts of guidance and control systems. Problems discussed are the organization of the guided missile department and the role of a special "reliability group;" meeting the requirements encountered by a missile throughout its life; the design limitations imposed by available components; test equipment requirements; and ground and preflight testing techniques.

SOMMAIRE

Cette note discute des problèmes variés devant être solutionnés durant l'étude, la réalisation, les essais au sol et le vol réussi des missiles. Elle accentue l'importance de la sécurité de fonctionnement, et porte sur l'étude et l'essai du missile lui-même, spécialement les parties électroniques des systèmes de contrôle et de gouverne. Les problèmes traités sont: l'organisation du département du missile à gouverne et le rôle d'un groupe spécial "sécurité de fonctionnement" réunissant les exigences rencontrées à travers la vie d'un missile; les limitations d'étude imposées par les organes disponibles; les exigences de l'équipement d'essais; les techniques d'essais au sol et en pré-vol.

1. INTRODUCTION

The more complex a problem, the more highly organized must be attempts towards a solution. This paper discusses the problem of successfully designing, manufacturing, and testing a guidance and control system for a reliable guided missile. It is not written to provide assistance to the design engineer in devising or choosing a missile guidance and control scheme. That topic is treated by other papers included in this volume.

Basically, this paper is directed to men who are responsible for an entire missile project which spreads across the boundaries of engineering, manufacturing, and testing. The intent is to cover many important aspects of the problem but to treat each in no more

detail than is needed to establish their proper perspective in the whole. Most of the topics discussed are "just common sense" and almost "self-evident." The achievement of a good basic guidance and control design does not guarantee ultimate success for the project. A carefully considered balance must be achieved between such factors as the design and its execution; the form of organization and the available qualified personnel; and the desire for new invention and its high cost in time and money.

Experience shows that a sufficiently high level of reliability of missile electronics is exceedingly difficult to attain and is rarely achieved. Strong emphasis is placed therefore on the problem of attaining reliability.

*Bendix Aviation Corp., Mishawaka, Indiana.

2. ORGANIZATION OF THE GUIDED MISSILE GROUP

a. Importance of Central Control and Supervision

First let us consider the organization required to develop and produce a successful missile. Because of the magnitude of the task involved, the job is often split up among several groups. When this is the case, special emphasis must be put on retaining in one central group the control and supervision of the various contributing organizations. Frequently there are alternate approaches to the solution of a problem neither of which has any obvious advantages over the other. Unless there is a strong central group that can make a decision and enforce a particular approach, valuable time is lost. Since time is money, money will also be wasted.

b. Responsibility - Triad of Design - Manufacturing - Quality

In the business of designing and manufacturing guided missiles, it is the function of the engineering department to create a design. The function of the manufacturing department is to execute this design. I feel it wise in the guided missile business to set up also a third department and to give it a stature comparable to the engineering and manufacturing departments. The third part of the triad is a quality department. Fig. 1 shows the three departments.

c. Responsibilities of Individual Departments

With the triad of engineering, manufacturing, and quality defined, let us examine the interrelationships. Reduced to simplest terms, we have:

(1) Engineering

- (a) Interprets performance requirements of the customer and translates these into a specified working design.
- (b) Develops the experimental models and translates their design into working drawings and specifications.

(2) Manufacturing

- (a) Develops tools and processes and acquires necessary materials.
- (b) Fabricates parts and assembles missile.

(3) Quality

- (a) Develops test and inspection procedures and provides all necessary test equipment.
- (b) Conducts all inspections and tests.
- (c) Monitors product quality and reliability.

The above outline of responsibility is shown in block diagram form in Fig. 1.

d. Perfection Under Difficulties; Need for Reliability

Guided missiles are perhaps the most complex flying mechanisms yet attempted by man. If these units are to work unmanned under extremely difficult environments, an extreme degree of perfection in the design

and its execution is required. A high degree of reliability is also required. It is not enough to have developed a guidance system capable of the desired performance and small enough to fit into the meager volume begrudgingly assigned it by the airframe designer. It is not enough that it be sufficiently light to come within the weight limitations imposed in order that sufficient fuel can be carried. Nor is it enough that it be easy to manufacture and test in spite of a terrible complexity. In addition to all the above virtues, the guidance unit has to work faithfully every time, not just part of the time. It must have a very high reliability. The achievement of a high reliability does not come by fiat and, to my knowledge, no organization has yet discovered any royal road to success in the attainment of missile reliability. The guidance system with all of its electronics is one of the most difficult parts of a missile to make reliable.

e. Two Broad Areas in Achieving High Reliability

For organized attack, the problem of attaining high reliability can be broken down into two broad areas:

- (1) Conceiving a design that is inherently reliable.
- (2) Building a device based on this design in a manner to enhance its inherent reliability.

The engineering department is basically responsible for the first and the manufacturing department for the second. The quality department must assist. It assists the engineering department by feeding back analyses of failure data to be used in design improvement; it assists the manufacturing department by continual inspections and by performing all required tests.

f. Reliability Is a Virtue

Consciousness of the importance of reliability might be considered as a virtue that needs special emphasis in order for a guided missile group to be successful. From the very outset of the project, the thinking of the triad of engineering-manufacturing-quality must always be "reliability oriented." It must be very clearly established and realized that the responsibility for achieving reliability is shared by everyone in the entire guided missile group. A significantly reliable missile can be attained only by requiring each man to think and act in terms of maximizing reliability.

g. Special Reliability Group

There are many times when the easy way is to ignore the demands imposed by strict adherence to reliability requirements. A special means of keeping the "reliability orientation" alive is needed. To do this, a special reliability group can be set up. Further reasoning justifying the use of a reliability group is that, while the responsibility for the attainment of reliability must be shared by the whole organization, actually attaining reliability is so important that a special group is useful to emphasize and assist in reaching a satisfactorily high goal of reliability.

The duties of such a group are described below.

- (1) It has the responsibility of constantly emphasizing the need for and the importance of high reliability.
- (2) It must monitor the reliability of the product and function as a central source of data on tests of the missile and its various parts.

- (3) It must continually analyze the test data and relay the findings to the operating department to facilitate elimination of the specific failures.

It is extremely important that the reliability group be started at the beginning of the project. There are several reasons for this. The important ones are: data on equipment failure can be obtained in a consistent manner; such data are available to the designers in retaining good parts of the design and eliminating the bad; and an early start can be made to impress the importance of reliability on the entire organization. Since a missile project of any size can very easily find itself confronted with large amounts of data on reliability, an early adoption of machine methods of data sorting and compilation is useful both to insure accuracy and to minimize expenditure of valuable manpower.

3. ENVIRONMENTAL REQUIREMENTS

a. Environmental Parameters Must Be Defined Early

Before design of a missile guidance system can make the transition from breadboard to prototype, the environmental parameters must be listed and ranges of values estimated. The following types of difficulties are likely to be encountered.

- (1) Temperature. Under operating conditions, changes may cause drift in circuit performance due to temperature coefficients of capacitors, resistors, inductors, etc.
- (2) Altitude. Possibility of arcing at high altitudes must be prevented by proper design.

- (3) Acceleration and deceleration (transient or constant) causes damage such as wire breakage and shorting of terminals.
- (4) Vibration. High accelerations damage vacuum tubes, relays, gyros, etc.
- (5) Humidity.

All of the above factors must be considered in the light of the effects of combinations of the above occurring simultaneously and in the further environment of the various likely service life phases of:

- (1) Shipment.
- (2) Storage.
- (3) Tactical stowage.
- (4) Handling and launching.
- (5) Flight.

The flight environment is the obvious one to which the attention of the designer is naturally attracted. Nevertheless, the other environments can be equally damaging. For example, the usual guidance system is extremely compact. Normally, the flight phase is so short that heating generated within the electronics is not a problem. However, sustained operation for many hours under test may require limiting the amount of testing at one time, a production inconvenience and expense. Similarly, vibrations encountered in flight are many octaves above the low frequency vibrations introduced by transportation by railroad or truck. With some types of shock mounts for electronic equipment, this low frequency vibration can cause severe difficulties.

b. Environmental and Life Test Evaluation

(1) Prototype Tests

As soon as a prototype of the guidance system is available it should be put through a complete series of environmental tests. The sooner these tests can be run the better since any troubles uncovered will require immediate changes in design; it is essential that this information be obtained long before expensive investments have been made in components or tooling.

During these very early tests as much performance information should be obtained as is possible. It is not enough to determine whether the design meets specifications. Specifications should be met with sufficient margin to take care of material and manufacturing tolerances. Information from these early tests will permit remedial steps to be taken while this can still be done reasonably economically, i.e., before production schedules must be met.

(2) Accelerated Time Tests

One type of knowledge the missile designer would like to have early is how well his design will stand up under repeated tests or prolonged storage. It is desirable to conduct "accelerated time" tests in an attempt to predict the probable effect of lengthy periods of storage or operational use in the field. In these accelerated tests, the usual approach is to substitute extremes of environment for time. Thus, for example, successive tests run alternately first under high heat and then under high humidity and which would take a week to run, might be the equivalent of several months' exposure to tropical conditions.

The validity of accelerated environmental tests has not been well established. For instance, the prolonged soaking at extremely

high temperatures of rubbers, lubricants, plastics, etc., helps only moderately in predicting ultimate survival of such materials. However, the technique may be used on a relative basis and comparative reliabilities may be determined between various possible choices of material by subjecting all to the identical test.

(3) Life Tests

Included in any environmental test program should be judicious life testing of missile parts and components. This applies particularly to any electromechanical devices normally subjected to wear and consequent degradation. Although the flight life of a missile may be only seconds or possibly minutes, it is necessary for missile devices to operate satisfactorily for many hundreds of hours during factory test and final check prior to launching.

Rotating equipment, such as motors, dynamotors, and inverters, must be designed to deliver full power during many repeated cycles of test. Switching devices, such as relays, should be life tested under loads closely simulating the switching loads. It is not necessary to conduct life tests until complete failure except in unusual circumstances. The life of a missile may be specified as 500 cycles of on-off operation of five-minute duration. Hence, the life test of 1500 cycles of five-minute duration gives a safety factor of three without the necessity of testing to failure. Moving contacts in potentiometers must be tested over the full range of contact positioning under the anticipated electrical load for similar duty cycles. These are a few examples of the kinds of devices that require life testing above the normal range of environmental tests.

(4) Combined Environmental Tests

The usual approach to environmental testing is to consider one parameter as the variable and to establish standard conditions for all others. Thus, for example, a missile would be tested under vibration over the frequency range of 20-500 cycles per second, while the temperature is 22°C, the humidity nominal, and the atmospheric pressure 760 mm of Hg. It is much more difficult to do, but it has been found that much valuable information can be derived by operating the missile under combined environmental conditions. Hence, while the missile is in an altitude-temperature-humidity chamber its altitude is changed from sea level to, say, 35,000 feet (760-180 mm of Hg.), its ambient temperature from 22° C to -55° C, and with essentially no humidity, at the same time the missile is subjected to vibration testing. Under the combined conditions, the climbing of the missile from sea level to high altitude is more faithfully simulated.

4. LIMITATIONS PLACED ON THE DESIGNER BY AVAILABLE COMPONENTS

a. Meeting a Short Design Schedule Forces Use of Available Components

Success of a guided missile designer depends on combining the skills and knowledge of a tremendous number of very specialized designers. This is particularly true in the case of the electronics designer who must make use of many components in his design. In a missile there may be as many as 300,000 components. The electronics part of the guidance and control system may typically have:

100 to 200 vacuum tubes in 10 to 15 different types,

900 to 1200 resistors in 200 to 225 different types,

300 to 600 capacitors in 100 to 125 different types,

75 to 100 potentiometers in 25 to 35 different types,

60 to 80 crystal diodes in 10 to 15 different types,

70 to 90 radio or intermediate frequency coils and chokes in 30 to 50 different types,

50 to 60 plugs and connectors in 15 to 20 different types (750 to 1000 individual connections),

40 to 50 transformers in 15 to 20 different types,

20 to 25 relays in 5 to 10 different types,

2 to 4 gyros in as many different types,

5500 to 6000 solder joints.

From a consideration of the total number involved, it becomes apparent that no large number of components may be developed specially. Basic electronic components are quite standardized, commercially available, usually reasonable in cost, and available to specifications. As a matter of economy both in time and money, the designer will find the most reasonable approach to be one wherein his initial selection is made from already available components.

b. Redundancy

With only a few exceptions, the limitations imposed on the missile designer by weight and size requirements prohibit the widespread use of redundancy in electronic circuits as a means of producing a more reliable system. Almost always the components in the various circuits are connected in series. During flight, the failure of a single component in this series chain is usually catastrophic.

One exception, however, where redundancy as a technique of improving reliability is available to the designer is in the application of relays. Where a relay contact forms the link in the signal circuit chain, the designer should select a multiple pole relay and parallel the contacts. If the probability of failure of a single contact is 0.01, two such contacts in parallel, under the most favorable circumstances, have a failure probability of 0.0001.

c. Overall Reliability Related to Component Reliability

With such a large number of components involved in a typical missile guidance and control design and the fact that failure of any one of many could be catastrophic to a flight, it is instructive to consider how the overall reliability relates to the various components. Let us consider a typical missile electronics design with a number of components as indicated above. A crude estimate of the overall reliability of an electronic system can be calculated as follows:

$$R_o = (q_1)^{n_1} \cdot (q_2)^{n_2} \dots (q_N)^{n_N}$$

where R_o = overall reliability

$$q_i = (1 - p_i)^{n_i}$$

p_i = probability of failure of i-th component

q_i = probability that the i-th component will not fail

n_i = number of i-th type components in series

N = number of different types of components in series

Fig. 2 shows the probability of system failure as a function of failure rates for groups of components, based on the above formula.

d. Component Evaluation Program

Eventually, the missile designer will arrive at a so-called approved list of components around which his design is based. However, to delay design awaiting the completion of this approved list would be to waste valuable time. There is always a base from which to start and a usually satisfactory approach is to build up this list as you go along. A few pointers may be in order on component evaluation.

- (1) Request early information on failures from the reliability group.
- (2) Use the evaluation process to compare equally available components to decide which component is the best.
- (3) Evaluate components on a priority basis by working on the poorest or potential trouble-makers first. Experience would assign to this category vacuum tubes and relays.
- (4) Evaluate components not only in regard to their electrical specifications but also in regard to the environmental specifications under which they must operate.
- (5) Establish a program of rating to reveal preferred suppliers (dependable and consistent quality).

e. Enlisting the Cooperation and Support of Component Suppliers

Among other things, the missile electronics designer is dependent for his success on his skill in making good use of the available components. A very positive step can be taken towards furthering his success by enlisting the cooperation and support of the various component suppliers. A reservoir of specialized knowledge and experience in coping with the problems of a particular component is usually to be found on the technical staffs of the better suppliers. By presenting their engineers with a statement of requirements and an explanation of their stringency, one can often challenge the engineers' pride with the result that they will outdo themselves to develop a more satisfactory component.

5. TEST EQUIPMENT REQUIREMENTS

a. Test Equipment

The facility with which a missile electronics designer can evaluate his designs is probably affected more by the availability of satisfactory test equipment than any other single factor. The electronics designer is dealing with complex variables such as error voltage output as a function of phase of a subcarrier amplitude modulation on a radio beam of microwave frequency. Hence, in addition to the more or less standard pieces of electronic test equipment such as oscilloscopes and signal generators, the designer is dependent on such special items as radar beam simulators and altitude simulators. Later in production test, perhaps both of these will be tied together with a programmer to command these special pieces of equipment to produce certain outputs at specified times.

b. Test Equipment Objectives

In test equipment design, it has long been the practice to attempt a precision ten times greater than the device being checked. Practical limitations of the state of the art often prevent attainment of this objective. If, for example, the missile electronics design is dealing with a high degree of stability in a particular oscillator and it is reaching the limits of the state of the art to attain this, it is not likely that a comparison test oscillator can be ten times better.

c. Test Equipment Design Started in Parallel with Missile Design

Even before the missile electronics block diagrams and instrumentation diagrams show some evidence of settling down, the test equipment design group must go to work. Ideally, from this time forward, the test equipment must progress in parallel with the missile design. Actually, at best, the test equipment can only keep up with the missile design. Certainly it is not practical for it to lead the missile design.

Design of test equipment is not as attractive as design work on the missile. Unless management is alert, it will find that test equipment is not given adequate attention and the entire program is delayed. As pointed out earlier in this paper, many organizations prefer to make the design of test equipment a responsibility of the quality department. This is because one of the basic responsibilities of the quality department is to perform testing of the missile. Speedy design and construction of test equipment is essential if the quality department is to fulfill its responsibilities.

d. Specialized Test Sets Often Worthwhile

It is usually possible to start with standard pieces of test equipment and to arrange and/or

modify them so that just about any special electronic tests can be conducted in the laboratory. The factory test requirements are a great deal different. In factory tests, it is necessary to provide specialized pieces of test equipment where more of the required intelligence is built into the equipment and less is required of the operator. Often it is worthwhile to start construction of specialized test equipment as soon as possible. This is particularly true when testing special functional circuitry.

6. GROUND AND PREFLIGHT TESTING TECHNIQUES

a. Testing Is An Important Part of Design

The cost of a missile flight test is high. There are many reasons for this. For one thing, no satisfactory way has yet been devised to flight-test supersonic missiles more than once. The greater the tendency to fail when in flight or the lower the reliability, the more expensive it becomes to get flight data. The flight sequence of a missile might be as follows:

- (1) Launch and boost to supersonic speed,
- (2) Stabilize in roll,
- (3) Lead into the guidance beam,
- (4) Acquire the target, and
- (5) Home on the target.

If failures are sustained in phases (1) through (4) in nine out of ten flights, then it takes ten missile flights to get one set of data on phase (5), the homing run. The cost of the homing data becomes impressive.

One method of reducing this high cost is exhaustive ground testing. Starting with all the missile electronics bits and pieces as the base of a pyramid, a system of successive 100 percent tests must be devised that has as its apex the preflight test prior to launch. Successive tests must be based on a satisfactory and conclusive prior test; the whole process must be carefully worked out so that nothing implicit in the planned sequence of events can do anything to invalidate the earlier tests. In other words, the testing process must remain under control. When inevitable troubles arise and replacements are made to eliminate the troubles, only satisfactorily tested parts of subassemblies or assemblies must be used in making the replacement. Just as the surgeon scrupulously scrubs before the operation in order not to introduce germs, so constant care must be exercised in order not to introduce untested elements to the complete unit. When dealing with as many as 2,000 parts, the above rule is not always easy to follow. The lack of timely availability of tested parts is perhaps the most common danger faced.

b. Specific Test Procedures Must Be Written

As early as possible, specific standardized test procedures should be written for all parts, subassemblies, assemblies, and final guidance and control assemblies. The objective is always to attain consistency over the testing methods which must be depended upon to demonstrate adequate performance on the ground. The specific procedures must be adhered to. It is not uncommon to find a technician who feels he knows a better way to run a particular phase of a test and habitually tries different ways rather than that prescribed in the standard test procedure. A "better way" is fine, but it must

be so recognized. It may very well happen that when someone else comes to run the test, very inconsistent results are obtained.

For many electronics designers, there is not much fun in writing test procedures, and there is a tendency to delay the day of writing them. One way of overcoming this trouble is by division of labor. It should be the responsibility of the quality department to create the test procedures under the supervision of the design engineers. The quality department will use these procedures in performing tests and so they have an incentive to get these procedures well established as early as possible. It is unreasonable to test in a way that does not correspond to the intentions of the designer. To avoid misunderstanding, the test procedure writer should submit all procedures for approval of the responsible designer. An additional point is the importance of keeping the procedures up to date. It is difficult to get them written anyway, and it is particularly hard to get them kept up to date. Only supervision's insistence will solve this problem.

c. Final Test Important and Should Be Sensitive

The final test should be given careful consideration. It is particularly important to make a careful study of the final test procedure to determine whether or not all conditions that could cause operational flight failure will be detected by this final series of tests. If, for example, pulse jitter in the guidance radar could cause the guidance unit to fail in performance, each guidance package had better be tested on the ground to ascertain that it is sufficiently insensitive to pulse jitter. A second example: even though the theory of dependence on previous tests in the test pyramid has already been

agreed to, there may be certain critical components that should be retested every time a final test is run. A third example: the test should be so devised that a relay used to effect changeover from beam-riding to homing is either implicitly or explicitly checked each time.

The desirability of an overall test simulating on the ground the entire sequence of events eventually taking place in the flight is obvious. There are certain compromises that must be made, however, particularly in the later production test phases. In launching and in flight, every missile sustains a certain amount of shock and vibration. The question is, should each missile be shock and vibration tested? A survey would find missile testing accomplished with and without. Probably no satisfactory general rules apply unless it is the rule, "take a conservative approach." If shock and vibration tests are consistently showing up trouble, it should be made certain that the shock and vibration levels are not excessive. Excessive shock and vibration levels during testing may easily occur because the methods involved are arbitrary at best. Having proved proper energy levels and troubles are still being shown, consider the shock and vibration tests as useful adjuncts to the other static tests in filtering out incipient troubles.

7. CONCLUSION

This paper discussed factors that can affect the success of a project to design, make, test, and satisfactorily fly a guided missile. It is hoped that consideration of the significance of each in the particular situation that confronts the reader may be helpful in increasing and speeding success of the project.

REFERENCES

1. Wilson, B. J., "Analyzing Missile Electric System Reliability," Trans. AIEE, 75, II, 1956, pp. 206-213.

GLOSSARY OF TERMS

- Breadboard The arrangement used to temporarily connect electronic circuits in order to determine how they work. Little or no physical resemblance to the final product exists in a breadboard.
- Component The smallest functional entity into which electronics equipment is divided. Examples of components are capacitors, fixed resistors, potentiometers.
- Prototype The first model that is both electrically and mechanically like the expected final product.
- Redundancy A technique of adding in parallel to a needed element, additional elements that continue to function should the original element fail.
- Reliability The probability that a device will function successfully under all environmental conditions in service.
- System Often in guided missiles a functional entity used to serve some particular purpose needed by the overall missile. Examples would be a guidance system that gives the missile directions to the target, a control system that takes action by making the missile flight obey the guidance signals, a propulsion system that provides the thrust for flight, etc.

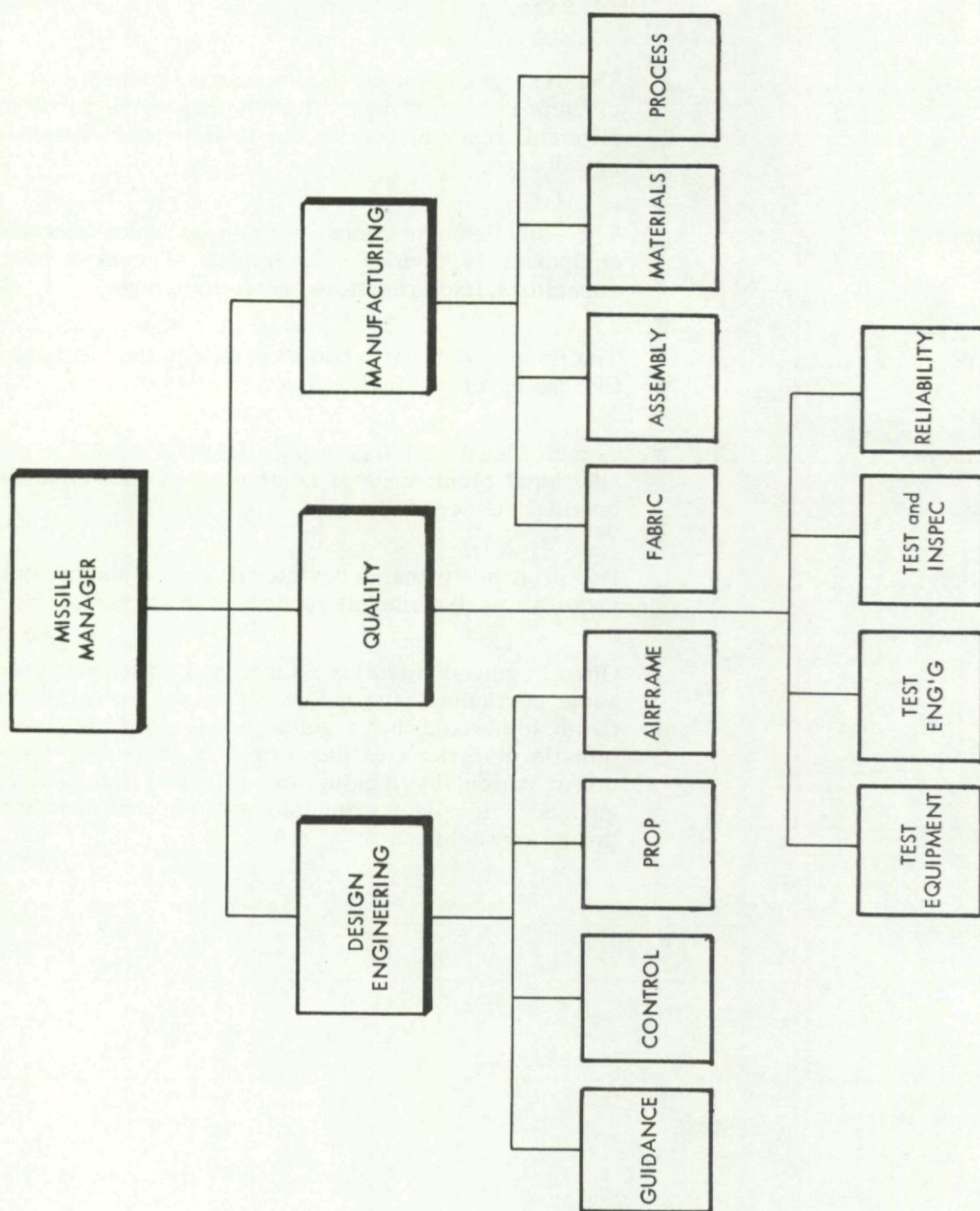


Fig. 1. The guided missile organization triad.

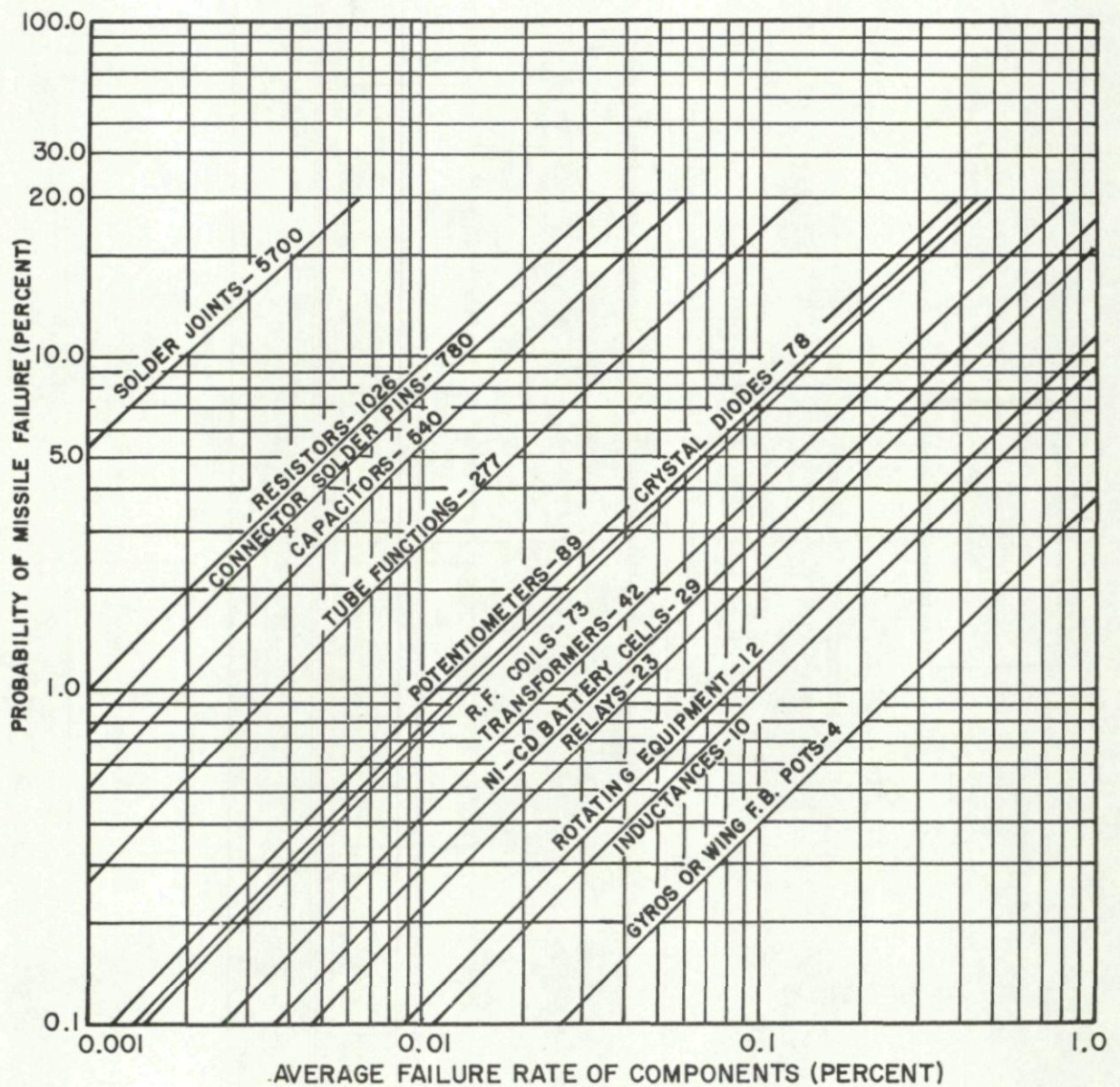
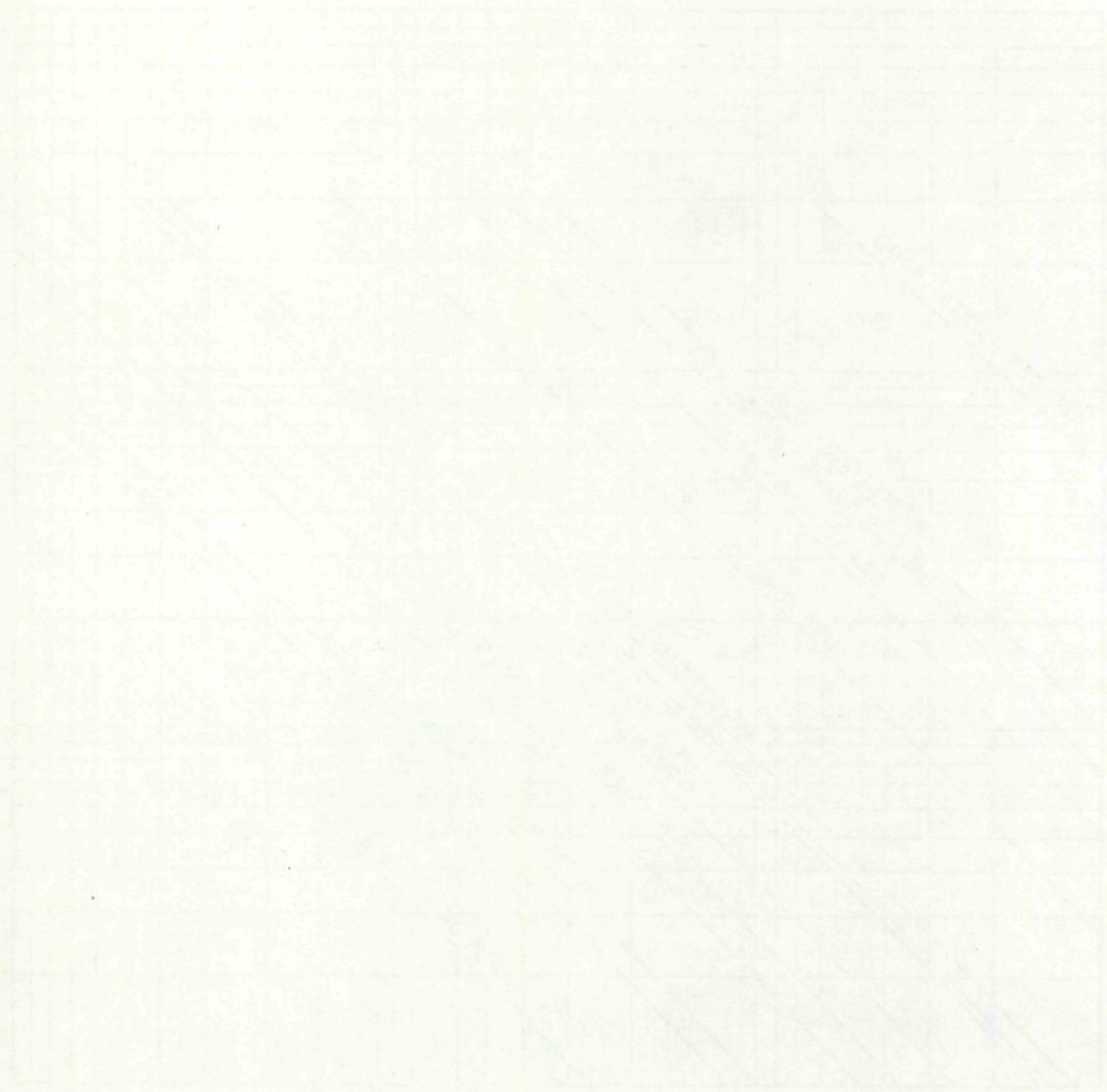


Fig. 2. Reliability requirements for specific electronic components.



APPLICATION OF METHODS OF SCIENCE
TO THE PROBLEM OF RELIABILITY
C. Raymond Knight*

SUMMARY

Modern statistical techniques permit quantitative study of reliability problems by the classic procedures of the scientific method: (1) Definition of the problem, (2) study of available data, (3) formulation of hypotheses, (4) testing of the hypotheses by experimentation, and (5) reevaluation of the problem in the light of results of the experiment. Application of the scientific method is illustrated in an experiment conducted by Aeronautical Radio, Inc., to test the hypothesis that maintenance procedures based primarily on tube testing adversely affect tube and system reliability, and further, that better reliability could be obtained through marginal-testing procedures applied to the equipment as a whole. Results of the experiment confirmed the theory that tube reliability is adversely affected by maintenance based primarily on tube testing; however, the tentative theory that system reliability is similarly affected was not confirmed in the particular experiment. The results also indicate that marginal-testing techniques, to be effective in reliability improvement, must be carefully chosen in relation to the characteristics of the specific equipments. The organized method of study exemplified in this experiment is essential to rapid progress in the solution of reliability problems in the field of missile guidance and control, as well as other areas of electronics.

SOMMAIRE

Les techniques modernes de la statistique permettent d'étudier quantitativement les problèmes de la sécurité de fonctionnement en suivant les démarches classiques de la méthode scientifique: (1) Définition du problème, (2) Examen des données disponibles, (3) Formulation des hypothèses, (4) Vérification des hypothèses par l'expérience et, (5) Nouvelle appréciation du problème à la lumière des résultats de l'expérience. L'application de la méthode scientifique est illustrée par une expérience faite par l'Aéronautical Radio Inc, pour vérifier l'hypothèse selon laquelle les procédés d'entretien basés essentiellement sur la vérification des lampes affectent défavorablement celles-ci et la sécurité de fonctionnement du système; et de plus pour vérifier qu'une meilleure sécurité de fonctionnement pouvait-être obtenue par des procédés de vérification marginale appliqués à l'équipement considéré comme un tout. Les résultats de l'expérience confirmèrent la théorie selon laquelle la sécurité de fonctionnement des lampes est défavorablement affectée par l'entretien basé essentiellement sur la vérification des lampes; cependant la théorie provisoire selon laquelle la sécurité de fonctionnement est affectée de la même manière ne fut pas confirmée dans cette même expérience. Les résultats montrent aussi que les techniques de vérification marginale, pour être efficaces et apporter une amélioration à la sécurité de fonctionnement, doivent être choisies soigneusement en relation avec les caractéristiques des équipements étudiés. Cette expérience donne un exemple de la méthode rationnelle d'étude qui est essentielle pour progresser dans la résolution des problèmes de sécurité de fonctionnement aussi bien en matière de gouverne et de contrôle des missiles, que dans d'autres domaines de l'électronique.

*Aeronautical Radio, Inc., Director, Reliability Research Department, Washington, D. C.

1. INTRODUCTION

The term "reliable" has been a part of our vocabulary for many centuries; and, as commonly used for qualitative or intuitive description, it has become widely understood. Webster's Unabridged Dictionary defines "reliable" as "Suitable or fit to be relied on. Worthy of dependence or reliance. Trustworthy."

This gross, intuitive meaning may provide an adequate measure when we are talking about reliability or trustworthiness of people, or even of inanimate objects, if we are willing to accept the status quo. If, on the other hand, we feel impelled to improve the reliability of these inanimate objects or these persons, and to do so with all possible speed, then a far more precise, quantitative measure of reliability is essential. "When you can measure what you are speaking about, and express it in numbers," said Lord Kelvin, "you know something about it." Real progress in science and engineering has always depended upon measurement performed in such a way that relationships between various factors may be expressed mathematically and that other experimenters may duplicate results through similar experiments.

We need not belabor the impelling need for rapid progress in reliability improvement in all fields of electronics, and especially in the field of missile guidance and control. It is the present urgency of this need that requires us to find ways of measuring in quantitative terms both the reliability that we have and the reliability that we must achieve. Measurement alone, however, is obviously not enough. An organized method of study of the quantitative effects is even more important.

The method of study that I shall recommend is not new; it is spelled out in the history of modern technological progress. I have

summed it up in the title of this paper: "Application of Methods of Science to the Problem of Reliability."

It may seem strange that I should devote this paper to a concept as elementary as the scientific method. Actually, my subject is the application of this old method with the aid of a remarkable new tool - modern statistics. Fig. 1 illustrates the basic elements of the scientific method.

Starting at the top of the cycle shown in Fig. 1, we have:

- (1) Complete, concise definition of the problem.
- (2) Investigation (accumulation and study of all facts and data available).
- (3) Generalization, in the form of tentative conclusions or hypotheses drawn from the specific facts and data.
- (4) Verification, through experimentation designed to test, and thus support or refute the generalizations.
- (5) Either repetition of the experiment under varied conditions or reevaluation of the problem or the generalizations, depending on whether the experimentation has supported or refuted the generalizations.

In essence, this method is a self-correcting process of inquiry and control, and one can readily see the analogy to the common servoloop.

My main interest, however, is in the fact that entire new areas of inquiry are now open to this well-established and traditional procedure. It is to point out that, with the new tools provided by modern statistical techniques, the scientific method can be applied with encouraging effectiveness to the solution of the problem of reliability.

During the early years of the growth of science, progress came most rapidly in the field of physical science. There were two important reasons for this situation. The first is that in the physical sciences, precise measurement, that is, measurement with small variability from one measurement to the next, could be achieved. The second reason is that, in the physical science laboratory, experiments could be carefully controlled to keep extraneous variables fixed and thus permit the effect of variation in a single factor to be reproducibly studied. In other fields of science, particularly biological science, early progress was hampered by the greater complexity of the experimental material and, in many cases, by the consequent inability of the experimenter to control extraneous experimental factors. It was largely in this area of biological science that the modern statistical methods were developed - methods which now permit quantitative, meaningful experimentation with an entirely new class of complex experimental material.

In short, during a 30-year span, the development of these statistical techniques has opened to scientific study entire new classes of problems - problems which, because of their complexity and practical impossibility of precise control, were earlier considered outside the scope of meaningful experiment. "Experiment," as the term is used here, is the verification type of experiment referred to in the fourth step of Fig. 1.

The significance of the new experimental techniques can be seen in their effect on the

study of reliability. In a reliability investigation, the fundamental "laboratory" is the farflung area where the product is being put to its intended end use. The task of controlling an experiment in such a "laboratory" would be virtually insurmountable were it not for the modern statistical methods mentioned. Particularly valuable are those techniques associated with experimental design and analysis of variance. In this paper, it would be impossible to describe these statistical tools in detail. The major purpose here is to generate sufficient interest in the possibilities of their application to encourage a wider use of the scientific method in studying and finding solutions to the problems of reliability.

2. DEFINITIONS LEADING TO QUANTITATIVE CONCEPTS OF RELIABILITY AND THE IMPORTANCE OF MEASUREMENT

The importance of quantitative methods in a scientific study has been stressed and it has been pointed out that the basic starting point for such a study is concise definition of the problem and of the terms to be used in the analysis of the problem. With these fundamentals in mind, definitions of several of the most important terms encountered in reliability work will be presented. These are definitions formulated with the intent that they shall be highly specific and usable for quantitative measurement. They are presented, of course, for the purpose of illustration rather than as an exhaustive list.

The first definition is of the term "reliability" itself. With minor variations, the following definition is becoming widely accepted: "Reliability is the probability that a product will give satisfactory performance for a given period of time under stated conditions."

The mathematical formulation of the definition is as follows:

$$R(X \in E; t) = P(a \leq X \leq b) = \int_E f(X; t) dx \quad (1)$$

where $X \triangleq (X_1, X_2 \dots X_k)$ = a vector of product performance characteristics

$C \triangleq$ "is contained in"

$E \triangleq$ the region of points defined as satisfactory performance

$a \triangleq (a_1, a_2 \dots a_k)$ } vectors specifying the upper and
 $b \triangleq (b_1, b_2 \dots b_k)$ } lower performance standards

$f(X; t) \triangleq$ the joint probability distribution function of the performance characteristics

$t \triangleq$ the length of the interval of observation.

Though probability itself is not directly measurable, it can be estimated from measurable quantities. Consequently, the above definition meets basic requirements. It is interesting to note that time, or more specifically, the time period during which performance is satisfactory is generally the quantity measured. Where the duration of a mission is relatively fixed, the proportion of total missions which are satisfactory (that is, the reliability) is often estimated directly. This definition of reliability, though much more specific than the dictionary definition, is nevertheless consistent with intuitive concepts of the term. Since probability is expressed on a scale of zero to one, and reliability by this definition is a probability, then reliability will also be expressed on a scale of zero to one.

Unreliability is merely the antithesis of reliability; it can be simply expressed as

$$U(t) = 1 - R(t) \quad (2)$$

where $U(t) \triangleq$ unreliability (a function of time)

$R(t) \triangleq$ reliability (a function of time).

This mathematical formulation is shown in functional notation as a reminder that both reliability and unreliability are intrinsically functions of time. Proposed definitions of two associated terms that are not so widely used but are noteworthy here, namely "maintainability" and "operational readiness," read as follows:

"Maintainability is the probability that when maintenance action is taken, a system will be restored to satisfactory operating condition within a given period of time;" and "operational readiness is the probability that the system will perform satisfactorily at any point in calendar time."

All of these definitions lean heavily upon the concept of satisfactory performance.

In the present state of missile development, it is practicable to monitor the missile trajectory with reasonable precision; consequently, in the estimation of missile reliability, a definition of satisfactory performance based upon the closest approach to the target has gained fairly wide acceptance. Thus, after the missile is considered ready and the final decision has been made to fire it, a clear definition of satisfactory performance can usually be formulated.

On the other hand, satisfactory performance as applied to missile operational readiness is not easily measured. A high degree of variability in the measurement of readiness is apparent when we realize that the decision to classify a missile as "ready to fire"

must be based on a relatively gross system-check at the prelaunch stage.* This final conclusive decision in effect is a composite of many individual decisions by a number of people, based on readings taken from various types of test equipment. The large degree of variability which results does not preclude meaningful measurement. It does, however, mean that many measurements must be made to permit the variability to be assessed. Experiments conducted to evaluate improvement must take this variability into proper consideration.

Operational readiness, while of only secondary interest during a development phase, becomes a primary facet of later missile utilization. As such, it demands greater attention than it has heretofore received. Most of the current figures on missile reliability are conditional upon the missile being in operational readiness; in other words, given a missile that is operationally ready to fire, it has a certain probability of successful flight to target. These reliability figures, or "conditional reliability" figures, thus are applicable only after the decision has been made that the missile is ready to fire. They take no account of whether three minutes or three days were required to ready the missile on the launching racks.

In the light of the definitions presented earlier, we now see that the product of the "conditional reliability" and operational readiness of a given missile becomes the probability that it can be fired at any particular time and that it will then come within a minimum prespecified distance of the target.

*Considerable study and discussion are being devoted to the determination of an "optimum" amount of complex test equipment for the measurement of operational readiness. (Stanley and Tampico, Associated Missile Products Corporation, Pomona, California.)

There are many more terms requiring definition in the study of reliability. The terms chosen for definition here have been selected to illustrate the types of measurement that are necessary if we are to make valid appraisals of improvement, or valid comparisons of different systems. In discussing the variability that must be reckoned with in these measurements, the intent has been to illustrate the need for and the importance of statistical tools. These tools and the electronic computer have made possible scientific study in fields where, a few generations ago, the physical scientist would have given up in complete despair.

It is not implied that the methods of measurement and the experimental techniques now available represent the ultimate in efficiency. Improvement is certainly possible, but we can nevertheless proceed upon the present basis. The primary gain to be expected from further refinement is a reduction in the number of observations required to reach valid conclusions.

3. RELIABILITY INVESTIGATION

Preliminary investigation and study of background material relating to the problem of interest is a fundamental step in the scientific method.

Most of the reliability information available today is best described as a large mass of observations, including: the relative frequency of component parts "failures," time-to-parts-"failures," time-between-equipment-"failures," and measured environmental conditions which have been considered severe or excessive - to mention a few. The importance of such investigational information, when cautiously used, must not be overlooked, for it is the only basis for good hypotheses. Publication of this material, within the restrictions of security, must be encouraged.

For an illustration of investigational findings of this type, reference is made to General Report No. 1 of Aeronautical Radio, Inc. (ARINC), dated January 4, 1954. In the investigation reported on, it was found that, on the average, one-third of the tubes removed from electronic systems in military service and returned to ARINC for analysis contained no discernible defect. The investigators at first assumed that, since tube defects vary according to type and application of the tubes, the characteristic defect patterns among tubes removed from similar equipments should be recognizable regardless of the location of the equipments. They observed, however, that the operating procedures and maintenance practices at any given military base mask these similarities; and that tube removals from a single type of equipment tend to be more analogous to the aggregate removals from all equipment types at the base than to removals from comparable equipments at other bases.

A portion of the data from which this tentative conclusion was drawn is shown in Figs. 2 and 3. These charts present comparative defect distributions of tube returns for similar types of equipment used at different bases and for all equipments under surveillance at each base. For this analysis, the tubes are grouped into four broad classes: those with mechanical, electrical, and miscellaneous defects, and those with no discernible defect.

Fig. 2 compares tube returns from fire control equipments at the Fort Bliss Army base and in Navy installations at Norfolk. The chart also compares these returns with the pattern of returns from all types of equipments at the respective bases. In Fig. 3, similar comparisons are made for tube returns from radar bombing systems at Carswell Air Force Base and at MacDill Air Force Base.

The greater similarity in returns from all equipments at one base as compared to returns from equipments of the same type at different bases is quite evident. Hence, the tentative conclusion that operating and maintenance procedures overshadow equipment similarity. It is also evident, however, that more than one hypothesis could be developed from the data in these charts. Moreover, any one or a combination of these hypotheses might be true in varying degrees.

In the remainder of this paper, the step-by-step procedures which were followed in the further study of this problem will be described.

4. GENERALIZATION AND PREDICTION

As a scientific generality develops, it is progressively labeled a hypothesis, a theory, and a law, depending upon how successfully and how broadly it has been verified. Whatever its state of development, any generality will encompass unobserved events - which includes all future events - as well as observed, or past, events. In effect, then, when we generalize, we predict. Thus the term "prediction" as applied to reliability is more than a mere catchword; there is justification in applying the term to any generality.

In the early phase of a scientific study, the investigator first assembles the available knowledge pertinent to the problem, then investigates the implications of the frequently apparently-unrelated observations. The next natural step is to look for general cause-and-effect relationships which "explain" the observations. This is the stage of hypothesis formulation.

A good hypothesis is one which can be put to experimental test on a quantitative and unambiguous basis. Such a hypothesis

can vary in nature from the simplest type, which will provide only a yes-or-no answer, to the highly sophisticated form wherein a mathematical model relating various factors in the experiment is tested. To illustrate the simpler type of hypothesis, the example previously introduced will be expanded.

As was shown in that example, early ARINC findings indicated that maintenance procedures and the experience level of maintenance personnel apparently have an important effect on electron tube and equipment reliability. On the basis of these findings, the following hypotheses were formulated:

Hypothesis 1. That, with complex equipment, it is difficult and often impossible for maintenance men of average technical background to identify the true source of trouble. More highly trained personnel would be expected to delve more deeply and accurately into the actual causes, and fewer tubes and other parts would be removed unnecessarily.

(The tendency to change tubes rather than delve into the true cause also has a psychological basis in that tubes are easy to change and are considered relatively cheap. For these reasons, there is a strong inclination to change tubes first and perform circuit investigation only as a last resort, in the event the tube change fails to "cure" the trouble.)

This hypothesis could be tested for a true or false answer by comparing reliability achieved on identical groups of equipments used in essentially the same type of operation but maintained, in one case, by normal military maintenance personnel, and, in the other case, by civilian technicians with engineering education and maintenance experience on the equipment.

Hypothesis 2. That maintenance procedures based primarily on the use of field tube testers have a detrimental effect on tube reliability and on system reliability; further, that preventive maintenance, based upon marginal testing procedures applied to the equipment as a whole, would give better tube- and system-reliability than would mass tube-testing.

This hypothesis could also be experimentally tested for a true-or-false answer by assigning different maintenance procedures to groups of identical equipments at a single location.

Both of these suggested experiments have been conducted. The analysis of the results of the first is not yet complete; I shall comment further on the second experiment in this paper.

An example of a more sophisticated hypothesis is currently being evaluated by ARINC in a new experiment recently initiated. A purpose of the experiment is to test the effectiveness of a miniature tube-shield insert in prolonging tube life by reducing bulb temperatures.

Increased temperature generally accelerates tube deterioration through chemical reactions, diffusion processes, desorption of gases, and other physical-chemical processes, the reaction rates for which are all reasonably well described by Arrhenius' Law. A mathematical model taking these factors into account has been assumed. This model hypothesizes that the removal rate of tubes will be equal to some constant independent of temperature plus another constant times the base of the natural logarithms to the power $-b/kt$; that is,

$$U = c_1 + c_2 e^{-b/kt}, \quad (3)$$

where b is the activation energy for the desorption of the gases from the envelope (believed to be approximately 0.16 electron volts), T is the temperature in degrees Kelvin, and k is Boltzmann's constant.

Enough experience is available on the electronic system concerned to permit assignment of approximate values to the constants. Thus, the change in tube removal rates due to the use of the shields is being predicted in advance for experimental verification.

Experiments of considerably greater complexity are being planned. One of particular interest will be designed to test a complex method of predicting reliability of entire systems, including consideration of such factors as primary supply voltage and regulation, component-parts temperatures, provision for ease of maintenance, and many other factors.

5. EXPERIMENT

Experimental test of hypotheses is generally more difficult in the field of reliability than in many areas of the physical sciences; in consequence, not enough effort has been devoted to experiment. We who are working in the field of reliability are unfortunately faced with the problem of developing the science at the same time that we are attempting to engineer reliable products. Since the function of engineering is to apply the knowledge of science, we have a dilemma. Reliability engineers must recognize the fact that they must develop the science as well as apply it, and they must have the support of management in their efforts toward these ends.

The danger of uncritically accepting generalizations based on investigation without repeated experimental verification seems almost a truism. Examples of this danger are legion in the annals of science. Indeed, in this home country of Galileo, it may seem an affront to consider this point anything other than self-evident. The mere complexity of the reliability problem, however, makes it perilously easy to imagine cause-and-effect relationships and then to accept them as scientific fact.

The following example of an experiment conducted by ARINC, with the cooperation of the United States Military Services, has been chosen for this presentation for two primary reasons, first, to illustrate how a reliability experiment can be carried out; and second, to demonstrate that negative results are sometimes just as important as positive results. It is important to recognize that we learn through making mistakes, even those mistakes which, in retrospect, seem obvious after the experiment is complete.

This experiment was designed to test the second of the two hypotheses stated earlier, namely, "that maintenance procedures based primarily on the use of field tube testers have a detrimental effect on tube reliability and on system reliability; further, that preventive maintenance, based upon marginal testing procedures applied to the equipment as a whole, would give better tube- and system-reliability than would mass tube-testing."

A group of 60 general-purpose military radio receivers were chosen as a suitable vehicle for the test. Pairs of these receivers are used in space diversity receiving systems at a fixed ground receiving site of an Army radio station. The 60 equipments

were divided into three groups of 20 equipments each. A distinct maintenance procedure was assigned to each group of 20 equipments. A brief description of each maintenance procedure is shown in Table 1.

With the exception of time required for preventive and emergency maintenance, all receivers in each group are operated on a 24 hour per day basis. This type of operation resulted in all receivers having very nearly the same operating time for the duration of the test, computed to be approximately 4370 operating hours. Records were kept on all reported malfunctions. The observation unit used in calculating the reliability function of an equipment, that is, the mathematical expression of its reliability as a function of a specified time interval, is the time between malfunctions other than those discovered at scheduled monthly or quarterly maintenance periods. Periods of time from an equipment malfunction to a scheduled maintenance and the time from the beginning of the test to first malfunction or scheduled maintenance

were included in the computations as incomplete observations. In addition to the records on the equipment, records were kept on the times at which electron tubes were removed, and on the number of maintenance man-hours required for each group of equipments.

In establishing the maintenance procedures to be used, several assumptions were made: (1) It was assumed that a characteristic of primary importance to the performance of these receivers was the sensitivity; (2) it was assumed that deterioration of the receiver sensitivity in time was primarily due to electrical deterioration of the tubes used in the equipment; and (3) it was assumed that reduction in electron-tube heater voltage would simulate a condition of deterioration to be expected later in life and, consequently, would tend to give a performance prediction. The results show that assumptions 1 and 3 do not appear to be justified. Nevertheless, valuable conclusions can still be drawn from the test.

Table 1. Three Types of Maintenance For Military Radio Receivers

Equipment Group Number*	Monthly Sensitivity Measurements: Criteria for Preventive Maintenance		Scheduled Quarterly Maintenance	Criteria For Tube Replacement
	At Normal Heater Voltage	At Heater Voltage Reduced by 10%		
I	When input signal needed for specified output is 4 microvolts or greater	None	All tubes tested on Hickok 539A tester	Below-minimum reading on Hickok-539A tester
II		When input signal needed for specified output is twice that needed at normal heater voltage	None	Adverse effect on receiver performance, as shown by tube substitution
III		None	None	

*20 equipments in each group.

The actual reliability functions estimated for each equipment group are shown in Fig. 4. In each of the Equipment Groups II and III, the reliability function fits an exponential curve very closely. The mean time to malfunction is 997 hours for Group II, and 969 hours for Group III. With the number of observations available, an appropriate statistical test does not show the difference between Groups II and III to be significant. The reliability function of the equipments in Group I does not fit the exponential distribution, thus indicating a possible difference in reliability between equipments in this group and the other two groups. Mean time between malfunctions, as estimated by non-parametric methods, is 1010 hours. Since this is not a single-parameter function, the mean alone does not adequately describe it; therefore, a comparison of means between Group I and Groups II and III is not an adequate comparison. An appropriate statistical test at 700 hours indicates a statistically significant difference in the reliability between Groups I and the combined Groups II and III, Group I having the higher reliability in the approximate ratio of 4 to 3.

It is of interest to examine the gross tube reliability and maintenance man-hours picture for these same equipment groups.

Preliminary data, based upon the first six months' surveillance, are presented in Table 2.

Here we find that over an equivalent time period, tube removals for Group I equipments are approximately twice the removals in either of the other groups. Similarly, maintenance man-hours for the Group I equipments are almost double the maintenance man-hours for either of the other two groups.

It may be concluded that, up to 700 hours, better reliability is obtained in equipments maintained as in Group I, but it is achieved at a substantial cost in tubes and maintenance man-hours. Weighing the advantages to be gained by the increased reliability against the cost of the tubes and maintenance time is a task beyond the scope of this presentation; however, it is a phase of the problem that presents an interesting and possibly important field for the operations analyst.

The results of the test also emphasize the economic justification for further work toward the development of more efficient means of marginal testing. It was determined during the course of the experiment that receiver sensitivity had relatively little correlation with the characteristics responsible

Table 2. Tube Removals and Maintenance Time for Military Radio Receivers, Under Three Types of Maintenance*

Equipment Group Number	Number of Equipments	Approx. Hours of Operation Per Equipment	Tube Removals		Maintenance Time	
			No. of Tubes	% of Total	No. of Hours	% of Total
I	20	4370	441	49.0	617.5	47.8
II	20	4370	233	25.9	333.0	25.8
III	20	4370	226	25.1	340.7	26.4

*Preliminary data based on first 6 months of test.

for most of the reported malfunctions. Receiver noise and, consequently, signal-to-noise ratio, seems to be a characteristic of considerably greater importance. It is believed that much of this noise may be caused by leakage paths forming on mica insulation in the tube spacers, but adequate methods of measuring this leakage, or the noise which it causes, are not now available. Repetition of an experiment of this type with revised marginal testing techniques appears desirable.

The results of the experiment definitely confirmed one portion of the hypothesis. This was the part which holds that maintenance based mainly upon tube testing adversely affects tube reliability. Other experiments conducted by ARINC also support this theory. On the other hand, the portion of the hypothesis dealing with equipment reliability was not verified, and, as a matter of fact, the result was negative. Our belief that revised marginal-testing techniques can be successful is nothing more than a new hypothesis which must also be subjected to experiment. I refer here to marginal-testing techniques which will give greater consideration to signal-noise characteristics of the receivers.

It is important to recognize the primary implications of this experiment:

First, it is too broad a generalization merely to state that marginal-testing procedures improve system reliability; the particular marginal-testing procedures for any specific equipment must be carefully chosen.

Second, there is not a one-to-one correspondence between apparent component reliability and system reliability; in other

words, it is possible to have an improvement in component-part reliability, with a net decrease in system reliability. This seeming paradox stems from the fact that definitions of satisfactory performance for systems and component parts are not necessarily consistent.

The first of these two primary implications may seem self-evident. Yet, the findings that have been presented seem to indicate that this obvious conclusion has not always been completely obvious. In the marginal-testing system which was evaluated in this experiment, the receiver characteristics tested were the same as those which had been checked for preventive maintenance purposes at the same location for many years. But the findings suggest that these characteristics had not been as carefully selected (for maximum effectiveness) as they might have been.

The second implication (that there is not a one-to-one correspondence between component reliability and system reliability) may also seem self-evident. But it now seems clear that an undue proportion of the blame for system unreliability has been placed on the electron tube simply because of a lack of appreciation of the truth of this statement.

* * * * *

The examples of reliability analysis-techniques presented in this paper were chosen with the hope that they would emphasize the importance of the organized method of study. Nowhere is the need for this scientific approach more compelling than in the field of missile guidance and control. Whether it be the missile field or any other area of electronics application, engineering intuition or hit-or-miss tactics

can no longer be solely depended upon if we are to make intelligently planned progress. At all stages of the approach to the problem of unreliability, the disciplines of the scientific method are essential. It is hoped

that the illustrations given in this paper will be suggestive of experiments that will not only aid in the solution of particular reliability problems but also add to the general store of scientific knowledge.

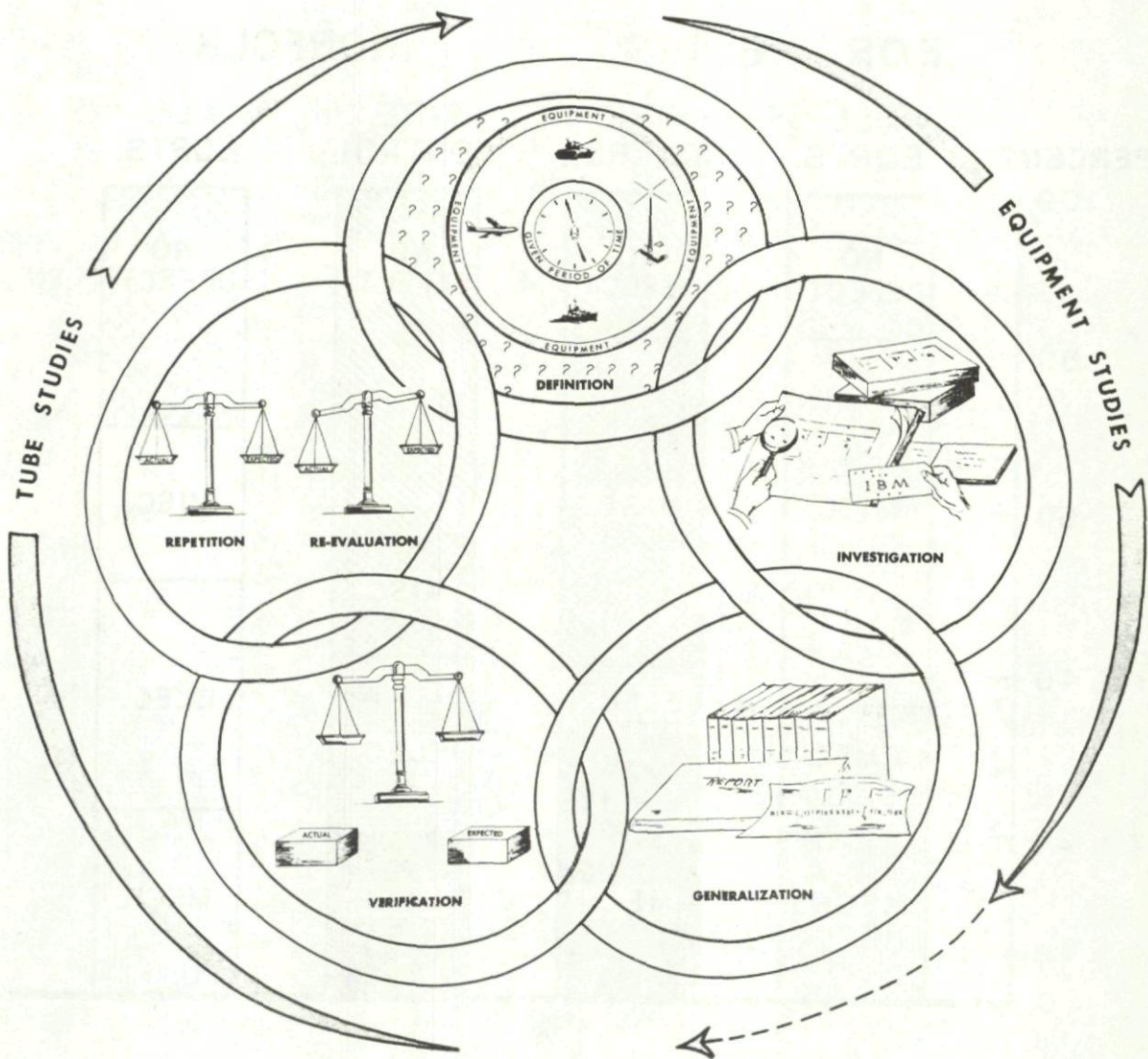


Fig. 1. Elements of the scientific method.

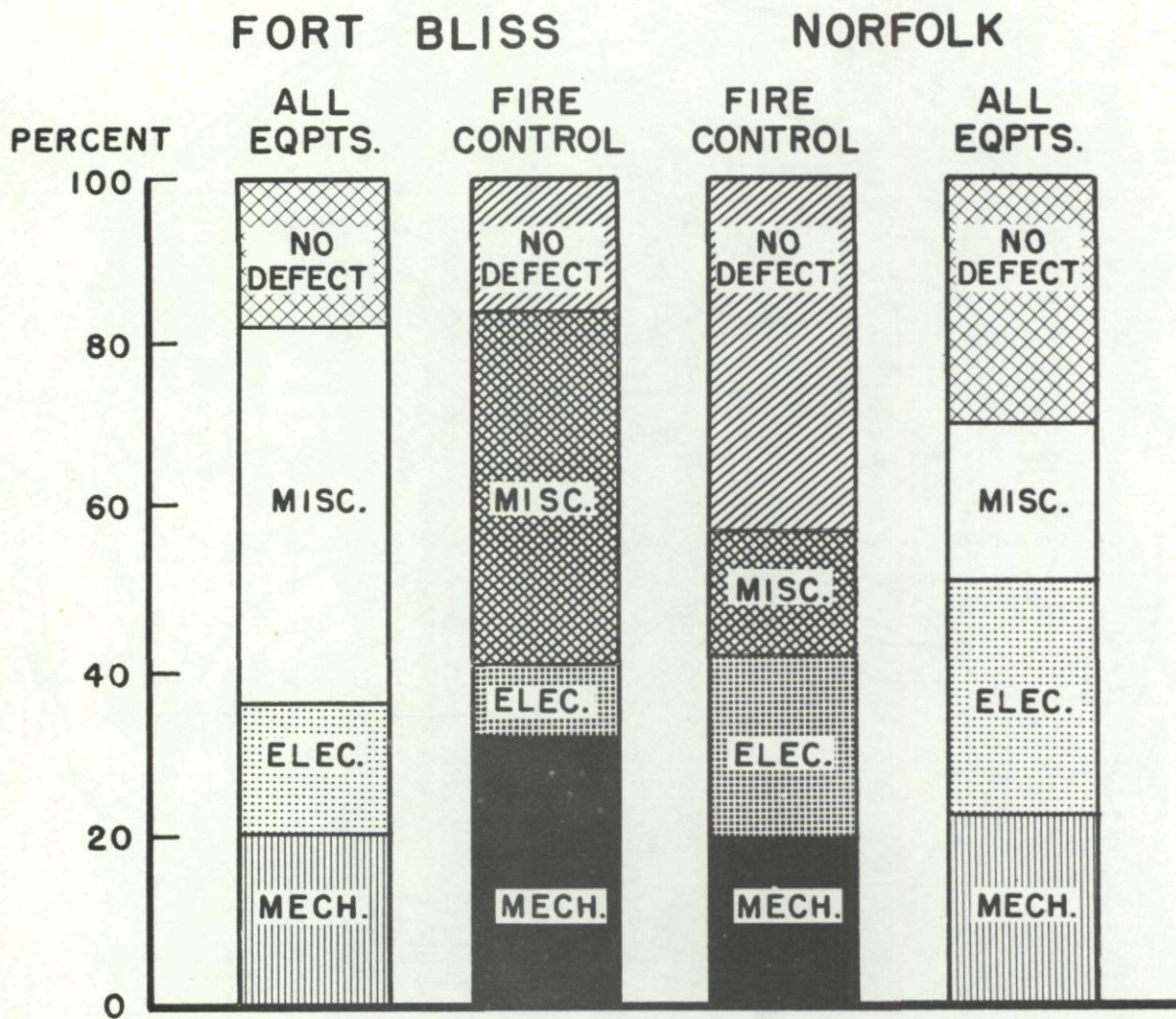


Fig. 2. Defect distribution of removed tubes:
Fire control system vs. all equipments at base.

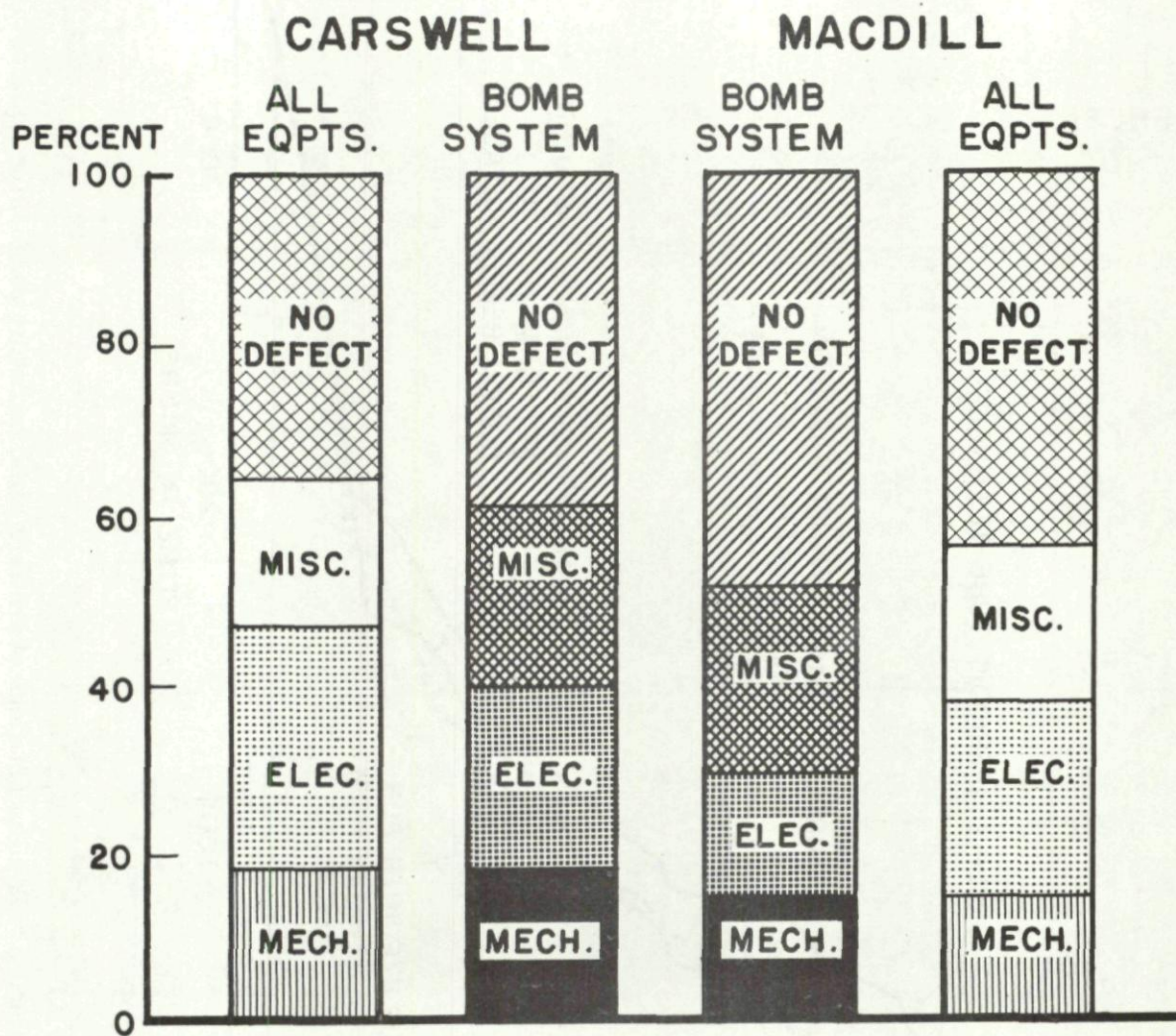


Fig. 3. Defect distribution of removed tubes:
Radar bombing system vs. all equipments at base.

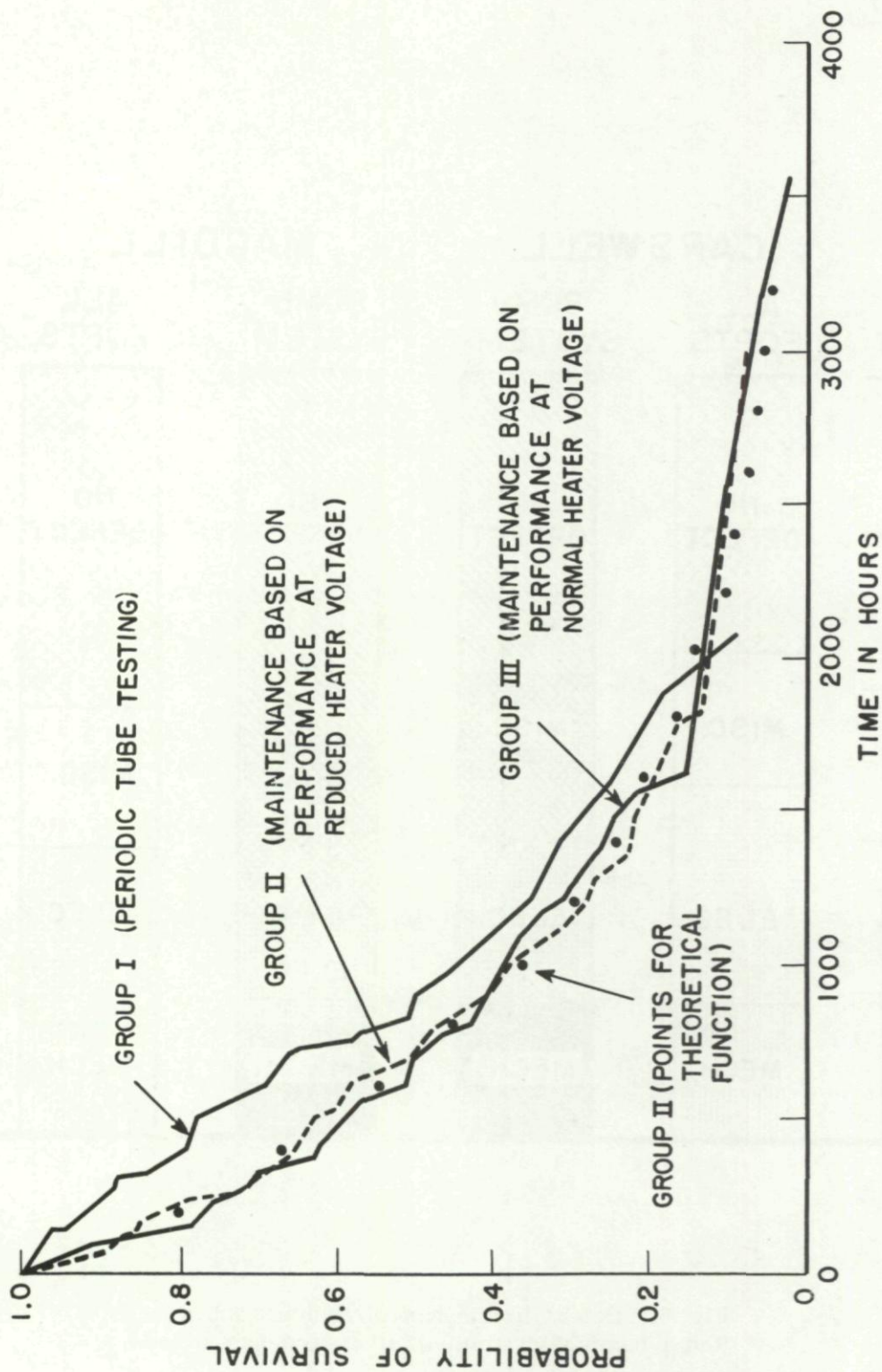


Fig. 4. Observed reliability functions of military radio receivers under three types of maintenance.

RELIABILITY OF GUIDED MISSILES

Edwin A. Speakman*

SUMMARY

In this discussion of guided missile reliability, special emphasis is placed on the importance of attaining reliable systems of guidance and control. With the advent of supersonic guided missiles whose guidance and control systems replace the human pilot and his attendant judgment, government and industry are faced with scientific problems never before encountered. The organizational pattern of the Department of Defense and the work it has promoted in industry to solve these important problems is described. In addition, certain specific areas are outlined where concentrated effort should yield progressive improvement. These recommendations are cited: (1) Reduce complexity, (2) devise means by which the adverse conditions encountered by guided missiles from development, through production, transportation, launching and actual flight can be anticipated and determined, and (3) develop planned programs of testing components to failure with emphasis on conditions of maximum stress.

SOMMAIRE

Dans cette discussion sur la sécurité de fonctionnement des missiles à gouverne, l'auteur dénote l'importance d'obtention de systèmes de gouverne et de contrôle dont le fonctionnement soit sûr. Avec la venue de missiles à gouverne supersoniques dont les systèmes de gouverne et de contrôle remplacent le pilote humain et son propre jugement, le gouvernement et l'industrie sont mis en face de problèmes scientifiques jamais rencontrés auparavant. Le modèle d'organisation du Ministère de la Défense et le travail qu'il a encouragé dans l'industrie pour résoudre ces importantes questions sont décrits. De plus, certains domaines particuliers où un effort concentré devrait amener une amélioration progressive sont décrits. Ces recommandations se dénombrent: (1) Réduire la complexité, (2) planifier les moyens par lesquels les conditions défavorables rencontrées par les missiles à gouverne du développement, à travers la production, le transport, le lancement et les conditions présentes de vol peuvent être anticipées et déterminées, (3) développer des programmes de planification concernant les essais des organes aux défections en se référant aux conditions d'accentuation maximum.

1. INTRODUCTION

With modern supersonic airplanes and missiles we are faced with scientific problems not heretofore encountered. New and radical designs in propulsion, aerodynamics

and guidance and control are required which have never been achieved in any military vehicle. Of special interest is the problem of guidance and control. In guided missiles, complex electronic devices must replace the human pilot and his power of reason and

*Vice President and General Manager, Fairchild Guided Missiles Division, Fairchild Engineering and Airplane Corporation.

control. If the guidance of a missile in flight is lost, it is very rarely regained and the whole purpose of operation including the missile is lost. It is, therefore, most appropriate that today we give attention to problems of reliability as related to guidance and control of missiles.

2. GENERAL SCOPE

Considerable progress has been made during the past few years improving the reliability generally of military electronic equipment. This has been brought about by study and continued improvement in electronic components such as electron tubes, electronic parts, and advances in equipment design, testing, and production methods. However, when the guided missile is examined, reliability takes on a new significance because the whole technology of missiles is novel and the function of missiles involves complex mechanical and electronic devices. These must operate under severe environmental conditions. So, while we recognize the enormous effort in which government and industry are engaged, the problem of solving these questions with particular reference to guided missiles remains with us.

In the following sections the reliability problem in general will be discussed with some reference to work done by the Department of Defense. Examples and data will be cited to illustrate those areas where concentrated effort will yield improvement. Since the guidance and control of guided missiles is primarily electronic in nature, most of the paper will apply to the electronic systems of guidance and control.

3. RELIABILITY GROUPS

In 1950 there was established an ad hoc group on reliability of electronic equipment under the Research and Development Board

of the Department of Defense. As vice-chairman of the Research and Development Board during the period from 1950 to 1952, the author organized this program and established the necessary policies for carrying on an active reliability program. At that time, though many agencies were aware of the reliability problem, it was found that coordination of these activities was necessary. (See Table 1.)

Table 1. Ad Hoc Group on Reliability of Electronic Equipment
Department of Defense

Task	Objective
1. Quantitative measurements	Establish minimum reliability figures
2. Development tests	Determine design criteria
3. Production tests	Verify reliability in production
4. Development procedures	Verify reliability in design
5. Specifications	Establish methods for specifying reliability
6. Procurement and contractual regulations	Determine compatibility with reliability objectives
7. Packaging and transportation	Avoid damage in shipment
8. Storage	Improve storage methods
9. Service exposure	Increase reliability by better maintenance methods

After intensive study the ad hoc group on reliability made several recommendations summarized as follows: (1) Failure data and performance data on all types of military equipment should be collected, summarized, and evaluated, (2) vigorous improvement programs for electron tubes and associated equipments should be continued, (3) procurement specifications of the military departments should include a reliability section, and (4) requirements as to the degree of reliability should be adopted in military specifications. While these are broad and general recommendations they serve as the framework for programs initiated by numerous military and industrial groups and provide a base for the comments that follow.

More recently this ad hoc group on reliability was reorganized and reestablished as the Advisory Group on Reliability of Electronic Equipment within the Department of Defense. Although it is somewhat premature to list the accomplishments of this group, and its working panels, it can be stated definitely that it has stimulated many industrial associations and individual companies to undertake increasing effort to solve the reliability programs with particular emphasis on guided missiles.

So much for the organizational background within the Department of Defense and the work it has promoted in industry. We consider the tasks which have been undertaken by this group.

4. RELIABILITY GROUP TASKS

The first complete program for the group has been fully developed and approved by the Assistant Secretary of Defense. Work is now currently underway in a number of task groups enumerated below:

(1) Develop minimum quantitative figures for reliability of the various types of military electronic equipment. The basis upon which the figures are determined shall

include factors of operational requirements, maintenance, complexity, and other significant factors.

(2) Develop basic requirements for tests to be accomplished on development models which will prove that the design is capable of meeting minimum acceptability reliability.

(3) Develop basic requirements for tests to be accomplished on production models which will prove that the equipment will meet the minimum figure for reliability.

(4) Investigate and recommend methods of specifying development procedures to insure that equipment designs will have the reliability required. Some factors are (a) reliability prediction, (b) component selection related to specific circuit and environment requirements, (c) adequate signal level, (d) effects of mechanical shock, vibration and temperature.

(5) Establish criteria and methods for specifying the reliability of component parts and tubes in terms of failure rate.

(6) Study present procurement and contracting practices and regulations to determine their compatibility with reliability.

(7) Investigate present practices of packaging for shipment and recommend improvements which will enhance reliability.

(8) Investigate the effects of storage and recommend improvements.

(9) Review present methods and procedures to assure that the reliability of equipment in service is kept up to the inherent design level. Factors included are: (a) maintenance based on performance measurements, (b) marginal testing, and (c) personnel training.

The interest evidenced in this program throughout industry has been very gratifying and helpful. The solution of the problems

involved will be difficult and will require continued close cooperation between the military and the electronic agencies.

Having given a picture of the Department of Defense organizational background for programs in reliability generally, attention is now directed to more practical aspects of the reliability program.

3. GUIDED MISSILES VS. AIRCRAFT

It has been stated that guided missiles are merely aircraft without a human pilot. This leads to the erroneous assumption that since human life is not involved, the missile need not be as reliable as aircraft. Just the opposite is true because the pilot in an airplane can adjust and substitute for malfunctioning of components, whereas in a guided missile all components of the guidance and control system must function without benefit of human assistance. They must therefore be completely reliable.

There are other factors which add emphasis to this conclusion. Most of the components of guidance and control are in series. Failure of any one single unit can cause the missile to miss the target and under these conditions it becomes impossible to recover, examine, or reuse the missile. It is obvious therefore that components in missiles must be much more reliable than those in aircraft.

6. DEFINITION OF RELIABILITY

But what is reliability? The Electronic Reliability Committee of the Radio Electronic Television Manufacturing Association defined it recently as follows:

"Reliability is the probability of a device performing its purpose adequately for the period of time intended under the operating conditions encountered."

7. V-2 ROCKETS

With this definition in mind, let us examine the record of a well-known German missile, the V-2 rocket. (See Table 2.) The data presented are based on tests made in the United States. In these tests, a total of 68, V-2 rockets were launched and of this number 3 percent were payload failures, 21 percent propulsion failures, and 29 percent guidance failures. It is interesting to note the large number of failures attributed to the guidance system. Of the total number launched, 47 percent worked correctly and 67 percent were usable.

In view of the history of the V-2 and the fact that some 3000 or more missiles had been used by the Germans in operations against England and Brussels, one might conclude that the tests described here are not impressive. On the other hand, we must assume that the V-2 missiles might have been affected adversely by shipping, handling, and general age at the time of launching. Information available, however, indicates that this record of reliability is not appreciably different from that in Germany.

8. COMPONENTS AND COMPLEXITY

The overall reliability of a guided missile or electronic system with particular reference to guidance is dependent upon the

Table 2. U. S. Tests of V-2 Rockets

Total number launched	68
Payload failures	3%
Propulsion failures	21%
Guidance failures	29%
Worked correctly	47%
Usable	67%

various components and units which make up the complete system. In particular, the complexity of a missile weighs very heavily in determining reliability. Complexity generally may be measured by the number of missile components which by individual failure could cause the failure of the complete guidance system.

These components may be classified generally as "series" components as distinguished from those components which if they failed would not necessarily cause the failure of the complete guidance system. These latter components are classified as "parallel" components. The failure of a single parallel component usually has little influence on the overall reliability. It is for this reason that we must concentrate our attention primarily on the series type components since they are most critical in their influence on the overall reliability of the guidance system.

9. EXAMPLES

Unlike more conventional systems, reliability of the missile guidance system depends not upon the average but rather the product of the reliability of each of the components. It is therefore obvious that the overall reliability of the missile decreases rapidly with an increase in the number of components.

For example, if a missile contains 100 components each having a reliability of 99 percent, the resulting reliability of the system would be only 36 percent. If a missile contains 1000 components each having a reliability of 99 percent, the resulting overall reliability of the system would be only .02 percent.

The formula upon which these observations are based is as follows:

$$\text{Overall Reliability} = R_1 \times R_2 \times R_3 \dots R_n.$$

Reference to this formula shows that if we should have 400 components in a missile, and this is not unusual, and a required reliability of 80 percent for the system, we could tolerate on the average not more than one failure in 1800 components.

While these are simplified examples, they do illustrate the parameters and enormous problems missile engineers have in designing reliable missile guidance and control systems.

In addition to these considerations much data have been accumulated based on the test and operation of military equipment in the field. This information is very pertinent to the subject of reliability. For example, comprehensive studies have been made and data analyzed to determine the impact and cost of equipments that do not operate. The Air Force reports that the cost of maintaining electronic equipment each year is about twice the original cost of the equipment. For the entire life of the equipment, it is estimated that the maintenance cost is ten times the original purchase cost. (See Fig. 1.)

In the Air Force one-third of the operating expense is for maintenance and one-third of their personnel is engaged in this activity, even though a large amount of maintenance is done by industrial contractors.

While these observations have reference to electronic equipments, generally, they are equally pertinent to missile guidance and control, which depends so much upon electronic components and systems. A prominent staff general in one of the military departments has stated that the availability of missiles for a military operation may be determined more by the availability of maintenance facilities and skills than by the mere number of missiles. Here, then, is an area where large dividends in the form of improved weapons can be realized by incorporating reliable components and designs.

10. MANUFACTURING RELIABILITY

With this background of data let us look further into the problem of reliability with special reference to manufacturing techniques. Components must be manufactured properly and have provision for 100, 200, or even 300 percent inspection. In view of the quantitative data cited above as applied to series components, it is obvious that a small percent defective components cannot be tolerated. We cannot afford to have a missile costing \$150,000 fail because of the failure of a 5 or 10 cent component. The primary concern is therefore in manufacturing processes; components cannot be made too reliable. A special statistical quality control system must be instituted for the manufacture of components to be employed in guided missiles.

11. DESIGN RELIABILITY

Of equal importance is the matter of design. A component may be fabricated exactly according to the designer's specifications but may fail when subjected to the conditions of flight. In such a case it is obvious that a component has not been properly designed. Design must take into consideration all of the environmental parameters. There is, however, a problem here which we must recognize. Design weaknesses are frequently determined by flight tests and these data are infrequently available because the missile is not recoverable. As a result, an unsatisfactory component may find its way into mass production before it has reached the required level of reliability.

Design reliability is therefore a unique and critical problem in the case of missiles. As a result of this situation, it is necessary to re-examine the environmental conditions and limiting test values. These may be carefully determined by scientific tests with unusually high safety factors evolved by numerous testing of units up to and including failure.

12. COST OF ATTAINING RELIABILITY

Based on my remarks so far it should be obvious that missile reliability cannot be evolved merely by trial and error over a period of time but must be achieved by a well-planned program. Such a program may cost a lot of money.

This question was examined by Dr. Erich Pieruschka of the Redstone Arsenal and the following resulted. Two guided missile programs were compared. In each of these programs, a total of one hundred good reliable missiles would be obtained. Program "A" was based on a planned expenditure of one million dollars, for reliability out of a total of two hundred sixty-eight million dollars. Program "B" contemplated an expenditure of six million dollars for reliability out of a total of one hundred forty-eight million. As illustrated in Fig. 2, Program "B" was by far the more preferable approach to the problem.

Recently, three Air Force jet trainers landed at an Air Force base in Texas within a few minutes of each other. General Holtner, pilot of the first aircraft landed without difficulty. General Schriever, pilot of the second airplane, had an electrical power failure and had to use his emergency system to get his landing gear down. General Davis, pilot of the third airplane, found that his landing gear indicator gave a faulty reading and he had to recheck in order to assure safe landing.

In commenting upon this operation, General Davis said, "Here were three of our most modern and reliable type planes and two of the three had electrical system trouble. If they had been missiles, two of them certainly would have been in trouble and probably failed completely."

General Davis commented further with regard to the importance of missile reliability and the extensive and complicated tests necessary to insure reliability. In one case which he described, a contractor had reported on the basis of laboratory component tests, that the missile would be 60 percent reliable, whereas actual flight tests showed a reliability of only 10 percent. General Davis explained that field and environmental conditions can affect the operation to such a degree that systems reliability can be decreased three or four times. He stressed the high cost of missile development but indicated that good reliability justified such expenditures.

13. CONCLUSIONS AND RECOMMENDATIONS

Some of the important aspects of reliability as applied to the guidance and control systems of guided missiles have been outlined here. The organizational pattern of the Department of Defense has been described briefly as well as certain industrial groups which

are contributing to the solution of this important problem. In addition, certain areas where concentrated effort should yield progressive improvements have been identified.

In conclusion, three recommendations are stressed: First, overall reliability decreases very rapidly with the increase in number of components. We must therefore reduce complexity. Second, to obtain reliability we must find some way to anticipate and determine the numerous adverse conditions which guided missiles encounter from development through production, transportation, launching and actual flight. Third, reliability of components can be improved by a planned program of testing to failure with emphasis on conditions of maximum stress.

While it is believed that missiles must become more reliable than we now know how to make them, it is felt that concentration on simplicity of design will provide major improvement. The Department of Defense and industry are working together as a team to solve the reliability problem and the solution will require the best effort that each can provide.

REFERENCES

1. Davis, Brig. Gen. Leighton I., 1955 American Institute of Electrical Engineers Technical Conference on Aircraft Electrical Applications.
2. Bridges, J. M., Director of Electronics, Office Assistant Secretary of Defense, Washington, D. C., "Electronics Equipment," p. 6, Sutton Publishing Co., New York, January 1956.
3. Clement, Lewis M., "Reliability of Military Electronic Systems and Equipments," Society of Automotive Engineers meeting, 21 April 1955.
4. Haviland, R. P., "Reliability in Guided Missiles," Jet Propulsion, Vol. 25, No. 7, pp. 321-325, 330, July 1955.
5. Martin, William H., Deputy Assistant Secretary of Defense, American Ordnance Association meeting, May 1955.
6. Lusser, Robert, "Reliability of Guided Missiles," Office Redstone Arsenal, Huntsville, Alabama, September 1954.
7. Pieruschka, Dr. Erich, "Optimum Allocation of Funds for Reliability of Guided Missiles," Redstone Arsenal, Huntsville, Alabama, January 1955.

INITIAL
COST

AVERAGE
MAINTENANCE COST

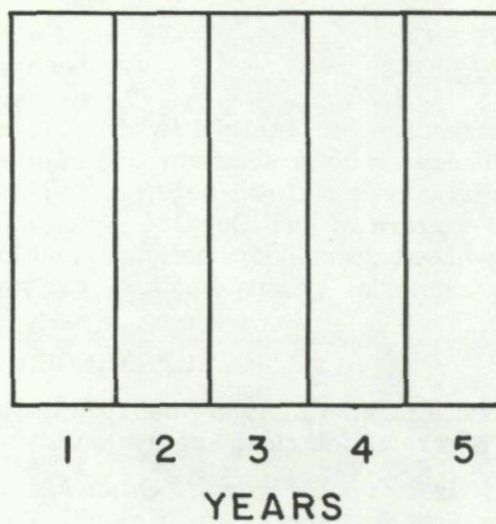


Fig. 1. Initial cost vs. average maintenance cost.

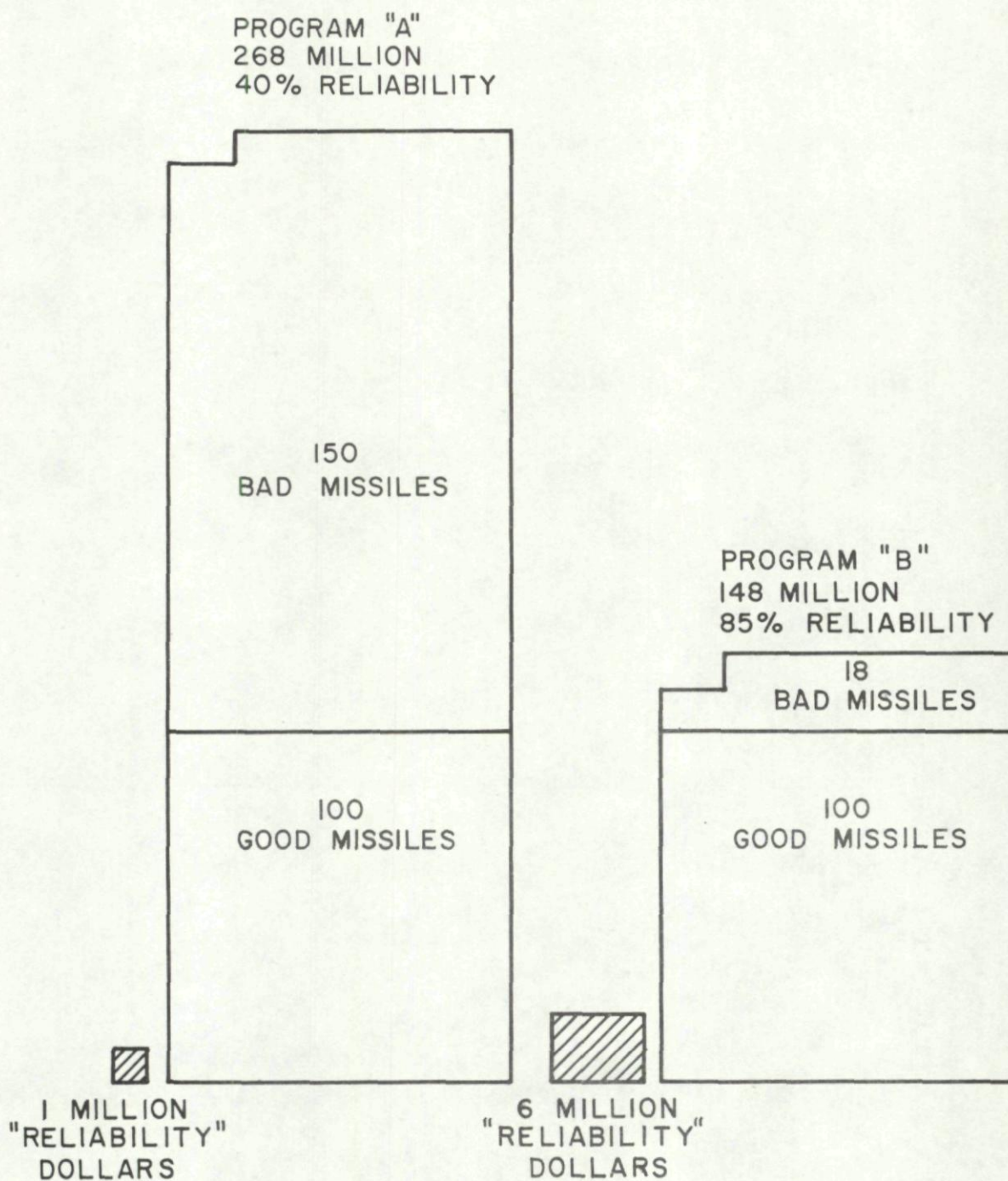


Fig. 2. Comparison between programs "A" and "B".

LABORATORY VS. FLIGHT EVALUATION OF AIRBORNE GUIDANCE COMPONENTS

Dr. W. H. Clohessy*

SUMMARY

This paper defines evaluation to include system limits, reliability, and effects of non-standard environment. It describes the organization and facilities of the Test Division and Electro-Mechanical Laboratories at the White Sands Proving Ground. It discusses factors favoring laboratory and flight evaluation and factors in the marginal area. It concludes with a discussion of current trends in this field.

SOMMAIRE

Cette note définit l'évaluation nécessaire pour inclure les limites du système, la sécurité de fonctionnement et les effets d'un milieu non-normal. Elle décrit l'organisation et les possibilités des Laboratoires d'Electro-mécanique du département d'essais du terrain d'épreuve de White Sands. Elle traite des facteurs favorisant l'évaluation en Laboratoire et en vol et des facteurs marginaux. Elle se termine par une discussion sur les évolutions courantes dans ce domaine.

The problem discussed here is pertinent not only to evaluation of airborne guidance components but also to the more general question of guided missile weapons systems as a whole. The nature of the problem is partly the apparent conflict between technical and economic considerations in such evaluation and partly how to maximize the information available at a fixed expenditure of men, money, materials, and time. In a very broad sense the problem implied in the title is not "which of the two is better?" but rather "what combination of the two modes of evaluation maximizes information for a fixed economic situation?"

The first step in evaluation is to define the purposes or objectives of the work. The second is to determine the nature of the system to be tested. This then allows the selection of the specific things to be learned or questions to be answered, and the variables

or environments to be imposed. From these can be determined tests to be performed, measurements to be taken, and calculations to be made. Execution of these tests and measurements leads to data which must be analyzed to obtain the answers to the questions. The above is idealized and assumes no variations in design or character of the system under test during the test period. In most cases this assumption is not valid and the whole picture is continuously under revision. The basic steps however remain essentially unchanged. They merely have to be reevaluated frequently.

If, at the point in the procedure where the "tests to be performed" and the "measurements to be taken" have been determined, we ask questions about accuracy, completeness of coverage, and confidence and if we invite the statisticians to prepare suitable numbers of trials and repetitions

*U. S. Army Ordnance Proving Ground, White Sands, New Mexico.

for our experiments, no one will be surprised at the economic infeasibility and the complete ridiculousness of the statistician's contributions. No country's national budget could stand the impact of such a blow.

We therefore impose additional restrictions, or rather, alternate restrictions. We drop the accuracy and confidence requirements and use instead numbers of missiles or components actually available and quantities of men, money, materials, and time as can be spared. It is here that the question of substitution of laboratory for flight occurs and it is here that we really dig in to try to plan an evaluation program which will give information in the quantity and of the quality meeting minimum requirements. Economics, in the broad sense, dictates our approach to the problem at hand.

A missile flight test is a relatively uncontrolled experiment. The in-flight environments are particularly difficult to determine and it is hard to assess their effects. The number of internal measurements which can be made is limited by the type and number of telemetry units which can be carried. Also, each component of the telemetry system is exposed to the same environment as the missile components and their behavior is hence modified by that environment in much the same way as that of the missile components. In addition, there is always the possibility that the installation and operation of the telemetry will affect the performance of the missile adversely. External measurements (electronic and optical tracking, etc.) are difficult to make reliably, and very expensive and they lead to large quantities of data to be processed and reduced. They do, however, give very precise values of trajectories, attitudes, and miss distances.

Measurements such as these may be of great use in the evaluation of specific components such as the airborne guidance units. This is true because the performance of a

guidance unit (other things being normal) determines trajectory, attitude, and miss distance. The number of values of environmental and performance variables for which flight tests can be made is greatly limited by the number of missiles to be flown. Even were it possible to fly sufficiently many missiles to cover all necessary values of such variables the task of sorting out the effects of individual ones on system performance would present the analyst with immense quantities of work. All of this costs large quantities of money, men, and time. But the greatest cost of a flight is the missile itself which, if recovered at all, serves only as a very poor source of information about the flight itself.

The bigger the missiles the fewer will be available for test. This is a simple economic fact. It is clear then that the big, long-range missiles with their immense requirements in terms of range instrumentation, their very precise components which should be monitored in flight, and their great cost make flight testing in large numbers prohibitive and unsatisfactory. Each flight must yield great quantities of information. But with a limited number of flights there are bound to be vast areas of variation of all the conditions and parameters which should be investigated, for a complete evaluation, which cannot be even touched upon. In addition to this, is the question of in-flight reliability which cannot be answered with any great confidence by a small number of flights. If the laboratory mode of testing missiles cannot be successfully exploited then we are doomed to "shooting in the dark" or to economic ruin.

Let us examine the possibilities for laboratory testing. Included under "laboratory testing" in this paper are component bench tests, component field tests, simulation, complete missile laboratory tests and complete system tests under laboratory (controlled) conditions. In short, we mean everything excepting only actual flights. If, during

a few flights, as much information about in-flight environments as is possible is obtained, then a component (say the guidance section) of the missile can be subjected at least approximately, to these environments many times in laboratory tests. Actual statistically valid data may be obtained and the component in question, factory rebuilt upon failure, to serve again and again in tests. The functional inputs can be simulated by analog computers and the outputs recorded or compared directly against recorded outputs of units in actual flight.

The effects of numerous conditions not experienced in flights but expected under battle use can be simulated and added to the normal input signals. The reproduction of such conditions in flight tests might pose almost insurmountable operational difficulties while their simulation in the laboratory might be fairly directly and simply achieved. Examples of such would be battle-field interference, countermeasures, or special terrain or weather features not present at the proving grounds. Reproducible inputs can be used to detect variations in output due to internal variability of the guidance section or to varying environmental conditions. Also, the number of breakdowns or malfunctions per unit operating time can be determined in statistically significant quantities. Tests to failure can be run to assess failure points and "safety margins" of either mechanical or electrical parts. Environmental and performance variables can be used over their entire ranges with any desired density of test points and at relatively small cost.

The total cost of equipment and personnel for all such tests for a period covering the entire evaluation of a weapons system may be equal only to from 2 to 100 times the cost of one of the missiles involved or perhaps one or two sets of ground equipment. The smaller figures apply to the larger missiles,

of course. This, added to the fact that information may be obtained which simply is not available in flights, makes laboratory testing a necessary supplement to flight testing.

While this picture of laboratory evaluation makes it apparent that flight tests can be made in limited numbers and the required information still obtained, great care should be exercised in the extrapolation of laboratory results to field conditions. Engineers and technicians in a laboratory characteristically treat the components under test with a healthy respect while troops in the field make repairs and settings according to standard rapid fire procedures. Random samples will not generally be used in laboratory tests and individual items may be repaired and reused frequently. Hence, the statistical validity of data from such tests may be questionable.

No amount of laboratory testing or simulation can yield really reliable miss distance data for anti-aircraft missiles. Also, the problem of reproducing accurately, in the laboratory, in-flight environments, such as shock and vibration, is a very difficult one. This, added to the difficulties of in-flight environmental measurements, particularly of vibrations with approximate noise spectra, makes necessary confirmatory flights to check out predicted component and missile reliabilities.

Finally, there are those whose principles do not permit them to believe in the results of anything but full scale field or flight tests of weapons. For them, flights against real targets are the only suitable means of evaluation. So we see that flights will never be done away with and the more flights the better is a fairly good rule of thumb.

I should like to describe briefly how the United States Army Ordnance Corps operates at White Sands Proving Ground in the evaluation of guided missile weapons systems and to

cover specifically the sort of laboratory and flight tests made there on airborne guidance components.

The center of activity in what is called the engineering evaluation of guided missile weapons systems at White Sands Proving Ground is a test planning and control staff. This consists of project engineers, technical specialists, and systems analysis people. They control the planning and execution of all tests and insure an integrated flight-laboratory program. The organization executing the flight tests, the Systems Test Division (STD), consists principally of military since it is our aim to test missiles under as nearly field troop conditions as possible. This division is organized by projects, by test teams of firing crews, assembly crews and a handful of civilian technicians for continuity.

The basic assignment of STD is the stockpile-to-target sequence of operations. Operational procedures for checkout, assembly, firing, and maintenance are strictly followed. When the missile is launched it becomes the problem of the White Sands Proving Ground range personnel to collect and reduce all tracking data and to transmit such data to the engineering test agency.

The White Sands Proving Ground missile laboratories, known as the Electro-Mechanical Laboratories, are four in number, consisting of a Guidance and Control Laboratory, primarily an electronics laboratory; a Flight Simulation Laboratory, composed largely of physicists and mathematicians using analog and digital computers; a Rocket Vehicle and Warhead Laboratory; and an Environmental and Instrumentation Laboratory. Plans from the staff are executed by one or both of the operating divisions and feeder reports are prepared there. These are consolidated and analyzed by the central

staff so as to insure maximum effectiveness and integration. In this way a coordinated test is achieved, well-balanced between the laboratory and flight aspects.

Given a specific weapons system to evaluate, White Sands Proving Ground technical personnel begin by examining the military characteristics desired by our field forces. In the case of the airborne guidance system, desired accuracy, reliability, and vulnerability to countermeasures or interference are the qualities which can be directly determined from the military specifications. Hence, our test program concentrates heavily in these areas.

A theoretical analysis of the system is made and all of the subassemblies and parts are subjected to scrutiny as to their ranges of operation, lifetimes expected, and environments anticipated. In some cases this analysis points directly to weak spots or predicted operational limitations. One or more guidance packages are put into our electronics laboratories for familiarization and operational tests under ambient conditions. At the same time, flights are begun with as much instrumentation as is possible. Instrumentation for shock and vibration and temperature environments as well as functional parameters (servo loop gain, etc.) results in records on airborne tape recorders or in telemetered data. The vibration data are digitized and put on IBM cards for processing in digital computers to determine the mathematical nature of the vibrations, i. e. for a given record the correlation functions and amplitude distribution curves, etc.

At this stage we have determined, in so far as possible, the in-flight environments and operating characteristics of the guidance package. Now, in a statistically designed set of laboratory tests we try to measure operational limitations and reliabilities under all environmental conditions considered to be

probable either in storage, in transportation, in checkout and assembly, or in flight of the guidance package. The data from such tests are analyzed by the statisticians and engineers and a probable figure is assigned to the overall reliability of the guidance section. Similarly, the limits within which the section performs its required functions are defined. If there appear to be operational limitations inside of the region of normal functioning, flights under those specific conditions will be performed to check out system performance.

Tests to failure will also be run in determining reliability for the following reason: If a given component operates too near the average level of failure (either in voltage, current, time, or other parameter) then the rate of random failures will be high and enough will occur during the above outlined tests to estimate the reliability under various environmental limitations. If, on the other hand, the level of operation is far below the average level of failure, the frequency of failure in normal operation will be low and a lower limit estimate of reliability can be made from the test-to-failure results directly. Examples of these two possibilities would be respectively a magnetron operating near maximum power output and the same magnetron operated at one-third that power.

Laboratory tests of this kind of guidance package take rather a long time, possibly a year to run a dozen units through sets of tests in various orders. At the same time,

however, flights will be continuing and other types of tests run in the laboratory. The effect, for example, of various tracking disturbances on an anti-aircraft target tracking radar can be determined in field tests and the subsequent system effects studied by simulation. The outputs of real or simulated computers can now be fed into the guidance package and its behavior analyzed.

Another type of problem which can be studied is the effect of various simulation signals of approximately the same frequency on the guidance package receiver. In this way the effects of countermeasures on the missile itself or friendly interference can be assessed. Such tests made during flights would yield only fragmentary information and might amount to very substantial field programs. In the laboratory, signal generators of all types are available and can be used in various combinations and sequences.

Naturally only a portion of what is possible in laboratory tests has been discussed here. It is believed, however, that it is clear that such tests supplement and in some cases actually replace flight tests. The number of flights possible grows smaller as the size and complexity of the missile systems increase. But the requirements for precise and extended test data grow in proportion to the size of the systems. Our only course of action in such circumstances is to make rapid and intelligent increases in the number and types of laboratory tests conducted.

TRENDS IN FIELD TESTING OF GUIDED MISSILES

Dr. Ernst A. Steinhoff*

SUMMARY

The paper presented describes the various efforts toward improvement of test range instrumentation and its adaptation to automatic data handling and analysis. The increasing availability of large analog and digital computing machines lends itself to mass handling of raw data as encountered in the data reduction and analysis field. To make use of this capability the output of the test range instrumentation must be adapted to direct input into the computers. Great efforts are presently being made to accomplish this and several projects concerned with real time data reduction and data analysis are in the development stage. Considerable savings in manpower in data handling and in overall cost of missile testing are expected. The manpower saved will be converted to computer programming and associated assignments leading to better availability of improved engineering information.

SOMMAIRE

La note présentée décrit les différentes tentatives vers une amélioration de l'appareillage permettant la série d'essais et son adaptation à la manipulation et à l'analyse automatique de données. La possibilité croissante des grandes calculatrices analogues et digitales contribue elle-même à la manipulation massive de données brutes comme cela se rencontre dans le domaine de l'analyse et de la réduction des données. En vue d'utiliser cette possibilité, la sortie de l'appareillage permettant la série d'essais doit être adaptée à l'entrée directe dans les calculatrices. De grands efforts sont actuellement faits dans ce but et plusieurs projets en rapport avec la réduction des données dans le temps réel et l'analyse des données sont en cours de développement. On espère réduire considérablement le coût d'essais des projectiles et l'énergie humaine nécessaire à la manipulation des données. La main d'oeuvre ainsi économisée sera utilisée pour la mise en programme de calculatrice et pour des travaux associés menant à une possibilité plus grande d'améliorer les connaissances dans ce domaine.

1. INTRODUCTION

In past years of flight testing of sophisticated guided missiles and pilotless aircraft, the necessity for more thorough preflight testing prior to actual launching has been demonstrated again and again. The increasing complexity of missiles frequently reduces the success chances of experimental firings in the early development phases to an extent

that a point of diminishing return is quickly reached. Recognizing this, testing techniques are being developed to reverse this trend and to reduce failure chances of complete systems to an economic level. In piloted aircraft a pilot can, in case of an emergency, change his flight plan and in many cases reduce or bypass a failing vital component. A pilotless aircraft or guided missile, however, does not have this alternative.

*Technical Director of Research and Development, Holloman Air Development Center.

The following paragraphs outline steps and approaches taken to increase the success of flight test missions and to reduce the number of misfirings of valuable test missiles.

2. PREFLIGHT ANALYSIS OF NEW MISSILE SYSTEMS

In order to arrive at realistic test plans, a complete dynamic preflight systems analysis is advisable to determine critical stability areas, shortcomings in aerodynamic properties and control surface efficacy, and other problems. This is done generally with analog computers. After completing this "dynamic survey," actual physical components of controls and guidance systems such as actuators, gear trains, airborne computers, control surface load simulators, etc., should be included in the open and closed loop analysis to study effects of nonlinear components and possible saturation of circuits on the overall dynamics of the system. After establishing the nonlinear patterns of components in closed and open loop analysis, these patterns can be simulated again in analog computers to facilitate the investigation of trends caused by the modification of performance and capabilities of individual components.

Complex interaction of components and component performance modification on the entire dynamic system can be studied and critical stability cases mapped for further exploration in flight tests. Since components frequently change their operating characteristics with changing environmental conditions, a thorough environmental test program must be set up and its effects upon overall dynamics analyzed.

3. COMPONENT RELIABILITY AND ENVIRONMENTAL TESTING

As outlined in the preceding paragraph, environmental conditions may change basic component properties and also may have

effects on the overall component reliability. Therefore components, assemblies, and even the complete system must be subjected to rigid and realistic environmental tests. Many of these tests can be performed by simulating environmental conditions in test laboratories.

Frequently, failures of components occur only when several critical environmental conditions occur simultaneously and not when components are tested independently under individual environmental tests. This indicates that realistic simulation of environmental conditions must take place. Testing to failure of the individual components should establish the actual operational limits of the component in question besides establishing the fact that the component meets the original specifications.

Since the individual components (see Fig. 1) must have a general reliability degree of 1 in 10^5 or better in a complex missile in order to result in a high degree of systems reliability, tests to failure must be extended to a high enough number of samples to establish this. This philosophy indicates the necessity of standardization of successful components to reduce the number of nonstandardized components on a missile to a tolerable minimum. Deviation from this rate is advisable only if this is mandatory to reach specified overall systems performance.

Results of reliability investigations showing the breakdown of various failure sources are indicated in Fig. 2. The diagram is typical for a great number of individual missile systems and shows that besides component failures, workmanship in assembly and human shortcomings in handling of missiles during the launching preparations also have an influence on the success of the overall mission.

4. CAPTIVE FLIGHT TEST RANGES AS TOOLS TO MORE REALISTIC FLIGHT SIMULATION

In the past decades, models as well as complete missiles have been tested in captive flights, carried by aircraft. Advances into the supersonic speed ranges make captive flight simulation more and more unrealistic since environmental conditions cannot easily be simulated closely enough, particularly if high performance missiles are concerned. Frequently the size of the missile involved does not permit airborne captive flight simulation.

In the last few years, a new tool, the high speed track or as we call it, the captive flight test range, has gained more and more popularity. On this track, with a future length of 10.65 km, sleds can be propelled far into the supersonic speed range and acceleration histories patterned after the actual inflight acceleration are simulated. Accelerations up to 57 g's and decelerations up to 100 g's have been reached. Full size missiles can be tested here under more realistic conditions than possible in captive free-flight testing.

One particularly attractive feature of the track testing is the recovery of the tested missile even in the case of a severe component failure. The sled design and the model or test item arrangement are important if the interference between the ground surface and the test item is to be avoided. In this case even the aerodynamic model tests can be performed with great success.

The captive high-speed flutter analysis of new airplanes and missiles is an important application of such a captive test range. Sleds are also used for linear acceleration tests of inertial guidance components and

as high-speed launching platforms for air-to-air missiles. The high-speed track as an aerodynamic tool is not a new idea. Professor Walchner, formerly at Goettingen, investigated and planned a high-speed track during World War II.

5. TEST RANGE REQUIREMENTS ON OPTICAL AND ELECTRONIC INSTRUMENTATION

The increase of speed and slant range of missiles and pilotless aircraft has doubtlessly increased performance requirements on optical and electronic test range instrumentation. Besides the improved tracking features and digitalized shaft output provisions, a semiautomatic tracking, in which human judgment is maintained, has proven to be most promising. To match the capabilities of the human readout of as little as 10 seconds, the automation, in the optical tracking field still must go a long way.

In the field of electronic tracking the situation is not very different. Considerable progress has been made by the application of chain radars which are controlled from one center and permitted to follow the missile throughout the various portions of its flight (see Fig. 4).

The changeover from one radar site to another is automatic and dictated by the quality of the received data. More powerful radars increase both the range of tracking even for small objects, and the tracking accuracy. Shaft outputs have been digitalized but they are not satisfactory in all cases.

Electronic triangulation type tracking systems like MIRAN (Fig. 5) have been developed so far that trajectory coordinates can be transmitted in a digitalized form and

then directly transferred into the electronic digital computers. From the experience of the MIRAN digitalization, the digitalization of Doppler data should not cause considerable difficulties. However, the digitalization should be performed after the position coordinates of the missile or aircraft are determined.

6. TENDENCY TO FLIGHT TEST ANALYSIS IN REAL TIME

The possibility of obtaining optical and electronic tracking data in digital form makes it possible to handle these data by large digital computers to obtain the flight path coordinates as function of time (location of object), i. e., velocity, and acceleration in terms of space-fixed coordinates, on a one-to-one time basis, or real time. Since telemeter data can also be digitalized, a transition from space-fixed coordinates to a missile-fixed coordinate system will permit the correlation of tracking and internal data with each other.

This possibility opens the road for a real time flight performance analysis. By introducing reduced tracking data, such as velocities and acceleration referred to the missile axis, into the missile equations of motion, flight performance data as well as dynamic response data, e. g., transfer functions and Fourier integral analysis data, can be obtained. This fact provides a completely new basis for the economy of flight testing, since the computer itself now can determine by the use of preset criteria, whether stability properties are as desired or must be improved. Signals from the computer to the missile can then change gain settings to improve or change stability parameters toward more favorable or more unfavorable conditions, the latter in order to determine critical stability conditions.

The economic aspect of such equipment is obvious, since the real time analysis yields stability data in a much shorter time. Additional modifications based on obtained data can be introduced during the same flight. Reduction up to 30 percent of the flight tests normally necessary is expected by the real time performance analysis. Several large U. S. proving grounds are converting their data reduction facilities for this purpose.

7. PROCESSING OF DATA

The processing of tracking and airborne data, semiautomatic thus far, can be made fully automatic if the tracking and internal data can be used jointly for reduction purposes. While external data are highly accurate for position information and in case of Doppler data also for velocity in the earth-fixed coordinate system, the acceleration information is mostly very unsatisfactory. If the accelerations in missile axis are determined by airborne accelerometers (accurate to one part in 10,000 or better) and then correlated with the external tracking data, highly detailed and accurate information can be obtained and the velocity data improved considerably.

The bottleneck in data reduction work is presently the reading and processing of raw data from either source. Particularly, tracking data require a considerable amount of smoothing by higher order polynomial techniques before further processing, while internal telemetered data present problems of noise removal affecting the quality of the data. To improve the quality, pulse width modulation and the pulse code type telemetry method are used, premitting higher accuracy, lower noise and, in case of the latter, on-missile digitalization of the signals. To compensate for poor geometry of tracking stations and poor performance of individual

operators, the data are over-determined by using more tracking stations than absolutely necessary, thus permitting the data processing computers to choose and use only the best combination of stations in each case. This procedure eliminates the tedious hand-reading of large quantities of tracking data.

The combination of analog computers with digital computers for data processing and data analysis reduces the overall requirement on computation equipment. The analog equipment furnishes derivatives of flight data instantaneously and also serves to furnish the first terms for iterations performed by the digital portion of the facility. The digital computer in turn monitors the accuracy of the analog computer to prevent drifting of the results in cases of amplifier drifts. Telemeter data, e. g., will be available in digital form for the digital computer, and at the same time in analog form for readout or printout purposes and visual display.

8. TREND IN TEST RANGE INSTRUMENTATION

The general development trend in test range instrumentation is toward self-contained automatically digitalized data production eliminating the need for hand processing or semiautomatic processing. Due to the simultaneous use of an excess number of stations to track the flying object, however, a human operator may monitor the data and also select the optimum combination of data at each given time. Human monitoring of those telemeter channels which contain the least noise and the most essential data is another variation in which human intelligence can still effect full automatic handling.

Particularly designed instrumentation radars having the required features and accuracy qualities are coming into being.

Besides continuous visual presentation of test results, magnetic tape records are maintained to store the initial raw data and the reduced data for an additional analysis if necessary. Since the handling of 60 or more channels simultaneously in one data reduction system does present, even for the largest computation facility, quite a strain, simultaneous tape recording permits later investigation of those phases which have not been included in the primary analysis.

The first and immediately handled data are of the quick-look type to indicate whether the flight mission was basically a success or failure. A real time data processing system normally handles between 30 and 50 percent of the incoming raw data only, since that number is needed to indicate the overall success or failure of a mission and to identify the overall area in which the failure occurred. Then a detailed analysis is performed to determine the details which caused a particular malfunction. In this way the available equipment can be flexibly employed.

The basic trajectory data acquisition is supplemented by long focal length optical systems to give necessary visual details and high frame rate cameras to permit slow motion interpretation of structural failures, of the separation of stages of multistage missiles, and of many other details. An interesting area is the constant improvement of optical and electronic equipment to measure miss distance for interceptor type missions.

Ballistic type cameras are not adaptable to the real time digital readout. However, comparators have been developed for the digital readout and they are being further improved. This information then can be

included in the second go-over in data reduction to supplement the real time information by more exact and accurate information. In order to obtain optical attitude information to serve as calibration for interval attitude reference, modifications on Askania Cinetheodolites have been proposed to permit the operator to rotate the crosshairs parallel to one of the main axes of the missile and digitalize the output of this rotation. Taking three instruments with this feature, a computer could compute the attitude of the missile and compare the results with the internal information. It is obvious that more detailed

information is obtained by the gyro reference, and that the effects of the gyro drift can be corrected by such optical attitude information.

In order to improve the optical acquisition of high flying missiles, chain radar equipment has been used to compute current acquisition coordinates. Several type optical tracking cameras, among them Askantias, have been adapted to this method; others will follow. Ballistic type cameras are triggered automatically as soon as the object reaches their field of view.

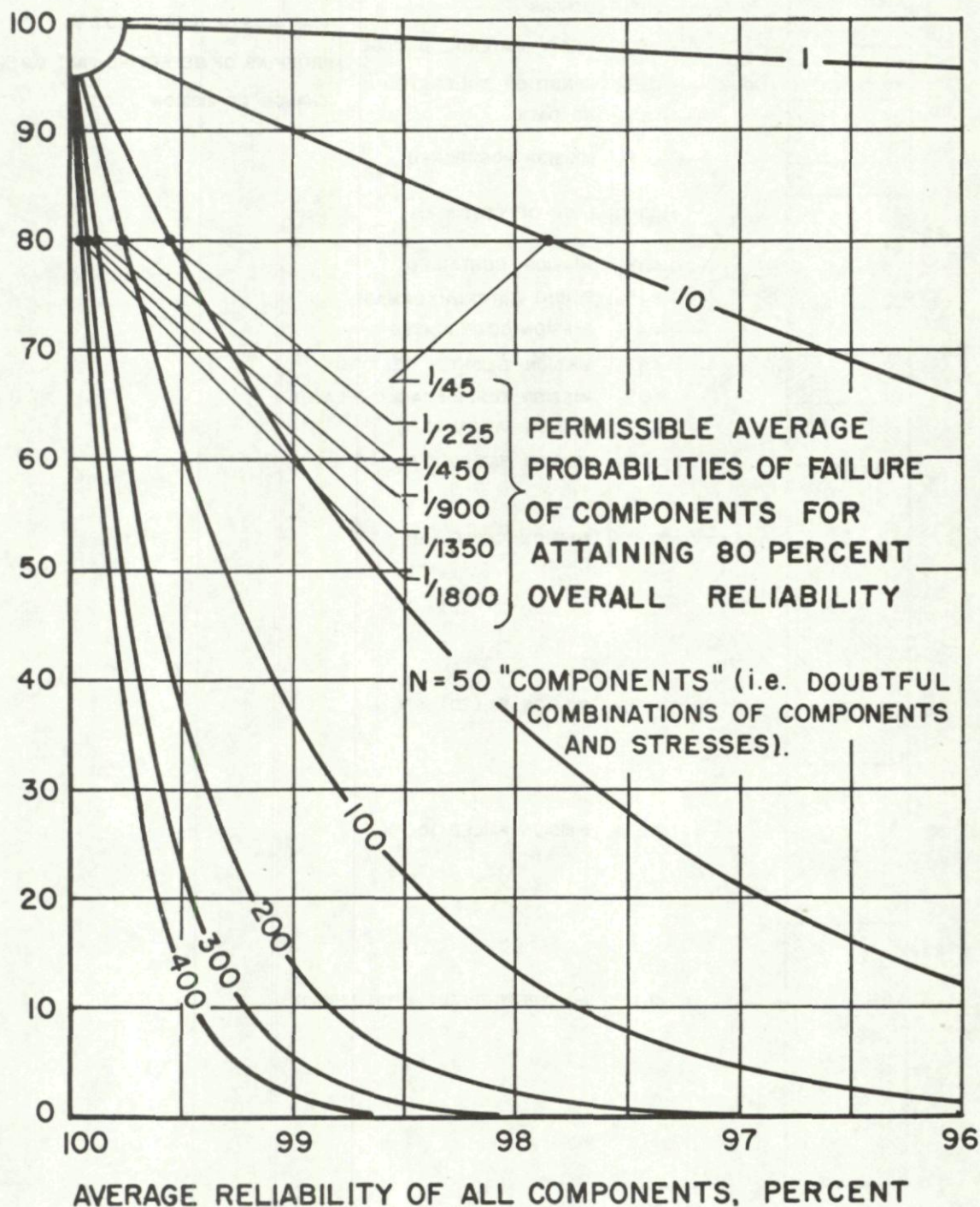


Fig. 1. Overall reliability as a function of complexity and reliability of components.

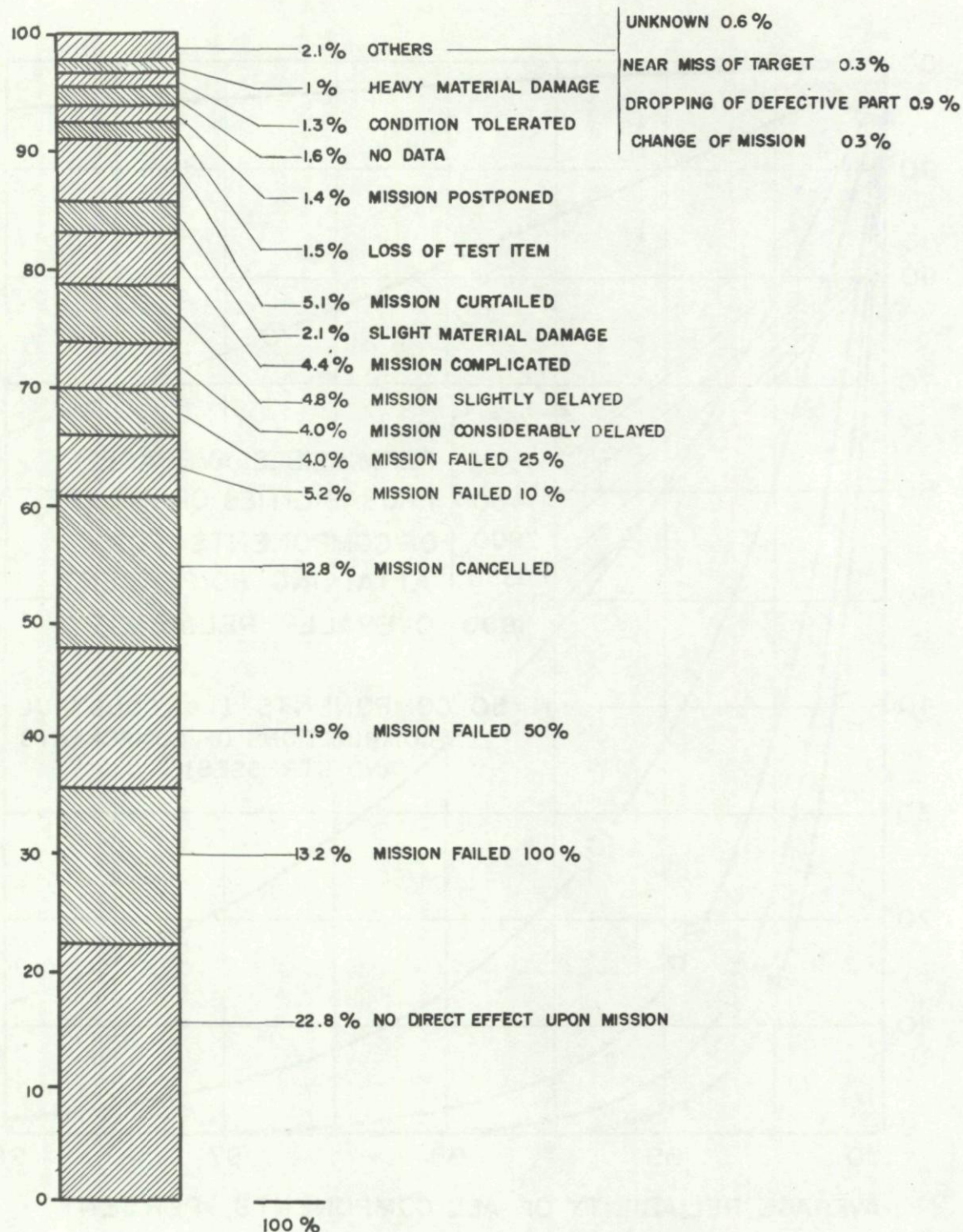


Fig. 2. Effect of failures (1200 failures).

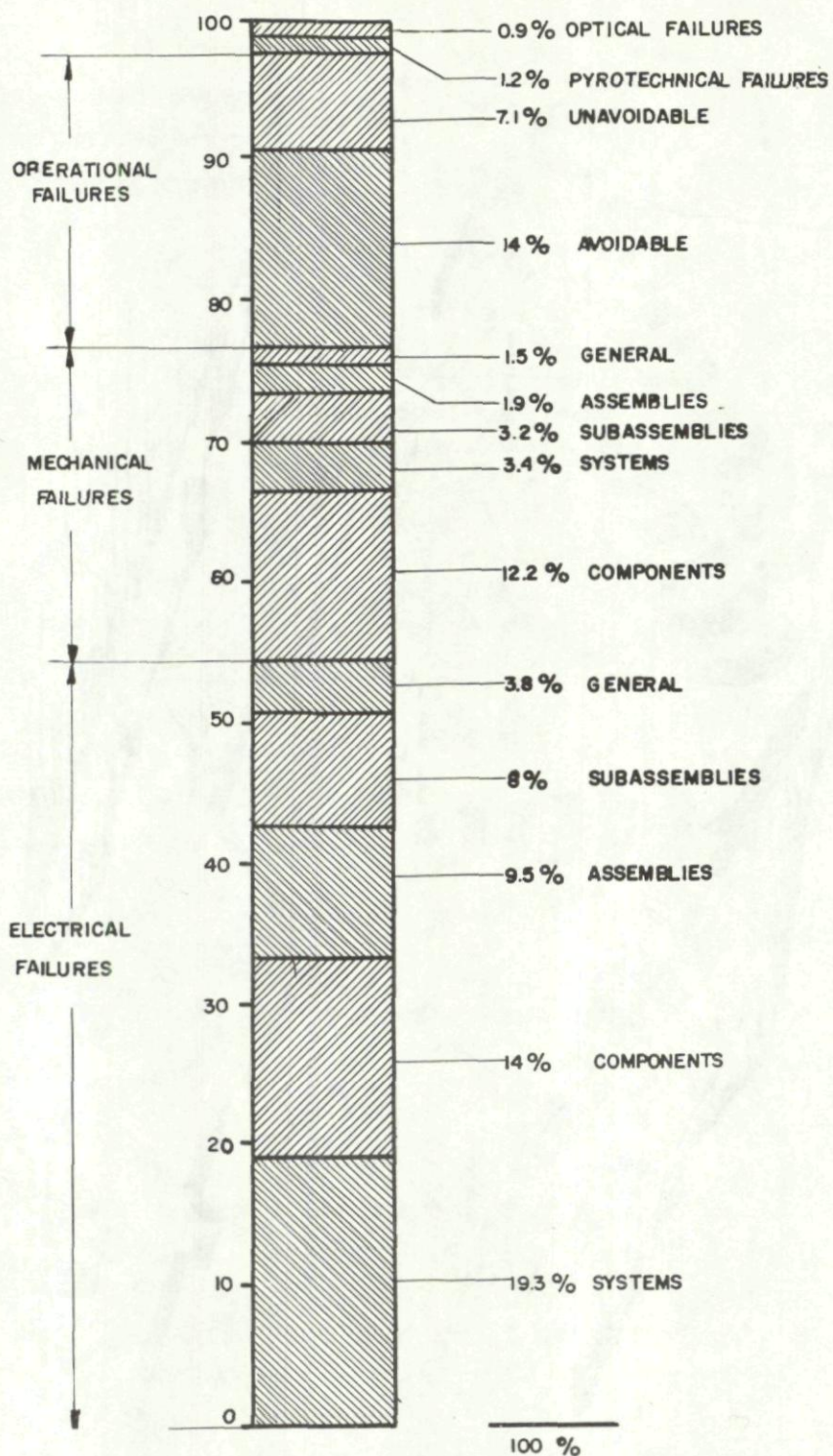


Fig. 3. Types of failures (1200 failures).

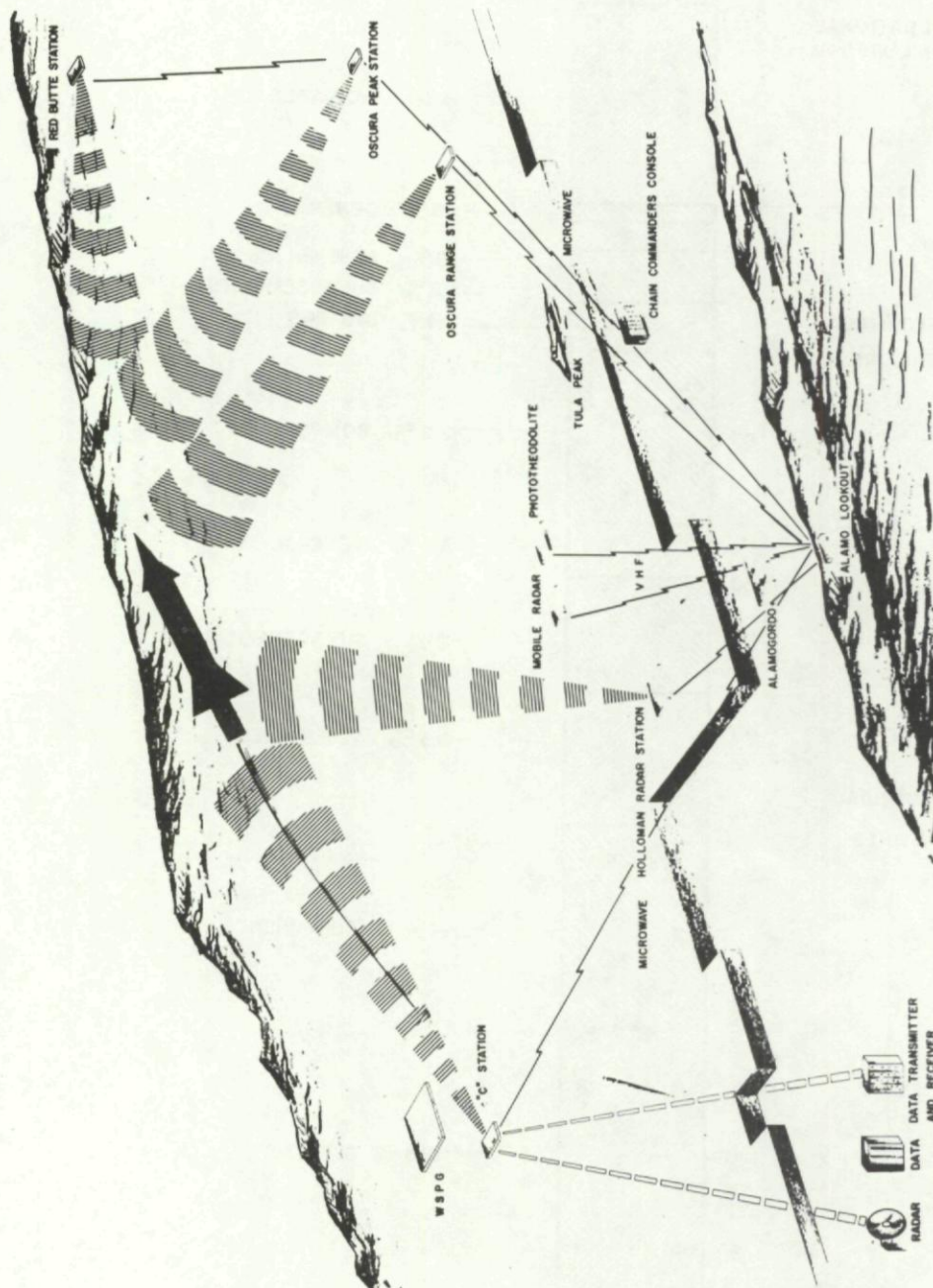
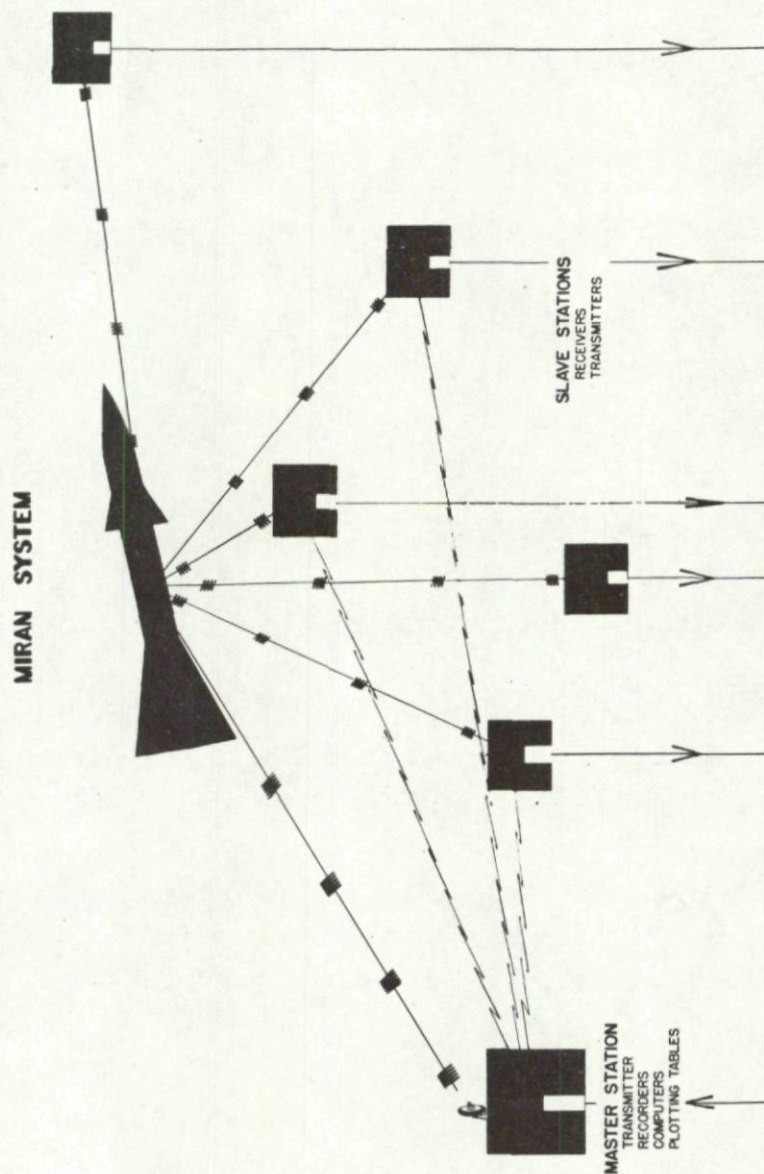


Fig. 4. Chain radar operation.



MASTER STATION TRANSMITS PULSES IN-MISSILE BEACON
 REPLIES ONLY WHEN INTERROGATED. SIGNAL RECEIVED BY
 AT LEAST 3 STATIONS, WHICH IN TURN GIVE PULSE TO
 MASTER STATION. REAL TIME EVALUATION POSSIBLE.

Fig. 5. MIRAN system.

LOW SIGNAL LEVEL MISSILE INSTRUMENTATION

L. G. deBey*

SUMMARY

During the past decade it has been practical to employ relatively unsophisticated radio transmission and receiving systems between missiles in flight and ground recording equipment in order to achieve satisfactory results from the trajectory instrumentation systems utilizing these transmission links. Recent trends in the guided missile program make it clear that future instrumentation systems will be required to operate at very low levels of radio signal energy with a consequent need for revision of the techniques which have been employed in the past. This paper discusses this general problem and suggests techniques which show promise of substantial improvement in instrumentation system performance under these anticipated adverse conditions.

SOMMAIRE

Durant ces derniers dix ans, pour obtenir des résultats satisfaisants à partir des systèmes d'instrumentation de trajectoire, il a été intéressant d'utiliser les liens relativement peu compliqués des systèmes de réception et de transmission par radio, entre les missiles en vol et l'équipement d'enregistrement du sol. De récentes évolutions dans le programme de missile gouverné indiquent clairement que les systèmes d'instrumentation futurs devront opérer à l'aide de signaux par radio dont le niveau énergétique sera très bas, ce qui impliquera par conséquent la nécessité d'une révision des techniques employées dans le passé. Cette note traite de ce problème général et suggère des techniques qui semblent vouloir promettre certaines améliorations des performances des systèmes d'instrumentation sous l'effet prévu de ces conditions défavorables.

The rise in technology following World War II pertaining to propulsion and guidance of missiles brought about a great expansion of effort in the instrumentation field. Measurement techniques applicable to the flight of artillery projectiles required a great deal of revision and extension to permit adequate collection of missile ballistic data. The past decade has seen a sizeable number of new types of instrumentation systems emerge from the research laboratories and these systems are currently playing an important part in the guided missile research and development

programs. In spite of the large effort expended in developing and improving these systems many of them are still a long way from being really first class technical solutions to the immediate instrumentation problems and fall far short of meeting imminent and more complex requirements.

Almost without exception instrumentation systems involve the establishment of a communication link between the missile in flight and one or more appropriate ground-based observing stations. The instrumentation

*Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland.

systems in use today employ more or less brute force methods to insure that the communication systems will yield an adequate signal from which the desired intelligence may be extracted. These statements are particularly true when applied to electronic instrumentation systems. In the use of radio communication links between missiles and ground stations there has been a trend toward the use of greater and greater radiated energy to compensate for the increased range capabilities of improved missile systems. Antenna systems of greater directivity and power gain are being used to further augment the received power. Such antennas require either detailed a priori knowledge of the flight path characteristics or complex auxiliary antenna-directing equipment.

There are obviously limits to the extent to which such methods can be employed and the effects of these limitations are clearly evident in instrumentation problems which are arising as a result of missile system technological advancement. One has only to scan casually the contents of a large number of technical journals, popular magazines, and the newspaper to find a variety of startling claims pertaining to the capabilities of modern missile systems. It becomes evident that the geographical boundaries of the earth will soon become limitations to our test range facilities.

We are, in fact, entering the era of space and interplanetary travel. The milestone marking the beginning of this new age will be the successful launching of the small artificial satellites currently being groomed for flight. Here we have missiles of such range and performance as to tax completely the ingenuity of the instrumentation specialist. The first of these vehicles will be very small and years will probably elapse before significantly large payloads can be put into orbit. Thus, for perhaps a decade, communication

systems will be required to operate over extreme ranges and extended periods with very small permissible radiated powers.

The solutions to meet these and future requirements exist only in the form of theoretical proposals. Some of the techniques required are a step or two beyond the current state of the art. The Ballistic Research Laboratories of the U. S. Army Ordnance Corps have proposed a partial solution to the particular problem of establishing sufficiently adequate communication with a very small artificial earth satellite to satisfy the experimental requirements of an ionosphere research problem. This paper will deal with some of the technical aspects of this proposed system. It should become apparent that the philosophy and techniques to be proposed would be equally applicable to the problem of instrumenting very long range ballistic missiles.

The Ballistic Research Laboratories have for many years been proponents of the use of phase measuring techniques for electronic ballistic instrumentation where the utmost in precision is required. The DOVAP system (Doppler Velocity And Position) installed at White Sands Proving Ground in early 1946, has consistently turned in superior performance from both the reliability and accuracy viewpoints. Furthermore, its use as a routine ballistic instrumentation system has paid a dividend by demonstrating it is also useful as a research tool for upper atmosphere research. Specifically, the Doppler data from a large number of high altitude firings have been used to determine ionization densities of the atmosphere as a function of altitude.

The instrumentation concepts to be presented in this paper were conceived to meet the requirements imposed by the use of this same Doppler technique in the United States earth satellite program, Project Vanguard.

The Berning (Ref. 1) ionospheric experiment is being readied as one of the possible and desirable upper atmosphere experiments utilizing the satellite as a vehicle for placing a source of electromagnetic radiation of optimum wavelength above the most dense portion of the earth's ionosphere for a reasonably long period of time. Payload for the satellite stage has been limited to two pounds, requiring that the crystal controlled transmitter and power supply, needed as a source of radiation, be held within this weight limit. Furthermore, it is desirable that the transmitter be capable of operation over a sufficiently long period of time to yield a large number of observations.

In the belief that the satellite will remain in orbit for a reasonably long time, if placed in orbit at all, the ionosphere transmitter is being designed to operate continuously for six months at a frequency of 74 megacycles per second. For a number of reasons this frequency appears to be near optimum. Weight considerations and the efficiency of transistors dictate that the transmitter cannot deliver more than about one milliwatt of radio frequency energy to the satellite antenna system.

The Doppler ionosphere experiment depends for its success upon the measurement of the apparent radial velocity of the satellite with respect to an observing station on the earth's surface. It is obvious that such measurements can be made most precisely when the apparent velocity is highest, i. e., when the satellite is moving most nearly toward the observing station. It is unfortunate that this condition first exists at the time when the satellite rises above the radio horizon and later, during the same orbital pass, when it disappears below the opposite horizon. For an elliptical orbit of two hundred miles perigee and eight hundred

miles apogee the slant range to the missile on the horizon from an observing station on the earth varies between roughly fifteen hundred and three thousand miles.

The difficulty of the communication problem now becomes apparent. Indeed it is virtually impossible to detect the missile radiation at the horizons because of the high attenuation of the signal at grazing incidence with the earth's surface. Fig. 1 is a plot of the propagation attenuation vs. elevation angle. It is to be noted that the attenuation at low angles of elevation is significantly greater than for the case of free space propagation. Based on these propagation data and the fact that the ionosphere experiment will not yield significant data at the high satellite altitudes where the ionization density is very low, it has been concluded that no attempt should be made to detect the missile signal below elevation angles of about sixteen degrees nor at altitude in excess of 300 miles. The maximum slant range under these conditions will vary from about 500 miles for a 200-mile altitude to about 800 miles for a 300-mile altitude.

The minimum received signal power at the terminals of the assumed 9-db antenna will be 3.16×10^{-16} watts or -155 dbw. The maximum signal power, when the missile is directly overhead, at the lowest altitude, will be 5×10^{-15} watts or -143 dbw. Since it is necessary to receive signal continuously during the entire passage above elevation angles of 16 degrees an essentially omnidirectional antenna must be used. A dipole, one-quarter wavelength above ground, is suitable. The reduction in receiving antenna gain at an elevation angle of 16 degrees is approximately 6 db. Further, the maximum gain of a dipole is 2 db less than for the 9-db antenna assumed in obtaining the results in Fig. 1. Assuming a transmitter antenna pattern factor loss of 15 db, the minimum received power will be -155 dbw - 8 db - 15 db = -178 dbw = 1.58×10^{-18} watts.

Fig. 2, prepared from data published by the U. S. National Bureau of Standards, shows the average levels of noise to be encountered as a function of frequency. It is to be noted that at the proposed operating frequency of 74 mcps cosmic noise predominates and is about $0.1 \mu\text{v/m}$ per kilocycle of bandwidth. This, for a receiving system having a 50-ohm input, is equivalent to a noise power of 2×10^{-16} watts per kilocycle of bandwidth.

A rigorous analysis of the effect of antenna pattern on antenna terminal noise power has not been carried out, but based on information furnished by V. H. Goerke (Ref. 2) of the National Bureau of Standards, it is estimated that the effective noise capture area for a horizontal dipole will not differ greatly from that of a vertical whip such as was used to measure the cosmic noise levels. Thus, it may be expected that the effective noise power at the antenna terminals will be about 10 db above KTB or 4×10^{-17} watts per kilocycle of bandwidth. Referring now to the expected minimum signal antenna terminal power 1.58×10^{-18} watts, it is seen that for a receiving system of one kilocycle bandwidth the S_p/N_p ratio will be approximately 0.04 while for the maximum signal case the S_p/N_p ratio will be about 125. The S_v/N_v ratios will be about 0.2 and 11.2 respectively.

One more piece of information is needed before the receiver design may be undertaken. The purpose of the communication link is to transmit intelligence. In the ionosphere experiment the desired intelligence is the shift in received signal frequency as a function of total ionosphere ion content and local ionization density at the satellite. For a homogeneous transmission medium, the Doppler shift in frequency as a function of polar elevation angle, for a satellite passing directly over the receiving site, can be described as shown in Fig. 3. The maximum frequency is seen to be about 1800 cps. The

actual received signal shift may be greater than 1800 cps due to instability of the signal source in the missile. It has been estimated that such effects might cause frequency changes as great as ± 1200 cps. The sign of the Doppler shift may also be either positive or negative depending on whether the missile is traveling toward or away from the observing site. Thus the total frequency shift may become as large as ± 3000 cps. The maximum first derivative or rate of change of Doppler frequency (\dot{f}) approximately 60 cps^{-2} and the second derivative (\ddot{f}) is about 1.5 cps^{-3} .

The maximum first derivative occurs at a polar elevation angle of 90 degrees while the maximum second derivative occurs when the polar elevation angle is 88 degrees 15 minutes. This second derivative defines the least post-detection bandwidth which may be used to pass the intelligence in real time.

The time-bandwidth concept of information theory permits the substitution of time for bandwidth, or vice versa, in transmitting a given amount of intelligence. The minimum bandwidth of the transmission link is fixed by the amount of information to be transmitted in a given time. If information storage cannot be accomplished, information must be transmitted as rapidly as it is being generated. However, it is often possible to provide storage for the information in which case the rate of transmission can be reduced and the bandwidth of the system can be reduced proportionally. In any event the time-bandwidth product will be a constant for a given amount of information.

It will be shown that in this particular problem the intelligence can be handled by the system in real time, in spite of the narrow bandwidths required to achieve adequate signal-to-noise ratios.

The maximum \ddot{f} occurs at an elevation as measured at the earth's surface of 57 degrees and hence the system bandwidth must be greatest at this angle. The received power, taking into account the pattern factor of transmitting and receiving antennas is 2.82×10^{-17} watts. For a one kilocycle bandwidth this results in a S_p/N_p ratio of 0.71 or a $S_v/N_v = 0.84$. Data of high quality from a Doppler system of this type require that the output S_v/N_v be about 10/1. Higher S_v/N_v ratios are desirable but not absolutely necessary. To achieve a $S_v/N_v = 10$, the bandwidth of the system must be reduced by a factor of $(10/0.8)^2 = 156$ or to 6.4 cps. Since the minimum bandwidth required for maximum \ddot{f} is approximately 3 cps the 6.4 cps bandwidth provides an adequate margin of safety.

At the lowest elevation at which it is desired to detect the signal the value of \ddot{f} is about 0.033 cps^{-3} and a bandwidth of 0.07 cps will be required. The 1-kc S_v/N_v ratio at this elevation angle was shown previously to be 0.2. Thus to achieve a $S_v/N_v = 10$, the bandwidth must be reduced by $(10/0.2)^2 = 2500$ or to 0.4 cps. Again an adequate margin of bandwidth is available.

When the satellite is directly overhead at the lowest altitude the S_v/N_v ratio was shown to be about 11/1 for an assumed 1 kcps bandwidth. Since this is adequate further filtering is not required. The bandwidth prior to the tracking filter, however, is greater than 1 kcps so the filter will be depended upon to reduce the bandwidth to at least 1 kcps.

The critical characteristic of the receiving system has now been specified. Certain other characteristics are desirable to enable the system to yield data of the highest quality

and to make it applicable to other similar instrumentation problems. These characteristics will not be discussed in detail but will be presented briefly to show how they affect the system design.

The receiving system should be capable of operating as a part of either a noncoherent or a coherent Doppler system. For the coherent system two essentially identical receiving channels are required; one for the missile signal frequency and one for the ground reference signal. The phase characteristics of the two channels in particular must be essentially the same, especially over the input signal level range of 0.1 to 10,000 microvolts.

The output signal should be in a form suitable for transmission via either wire lines or radio links to a central magnetic tape recording facility. These requirements prohibit use of a system in which the signal frequencies are at or near DC.

The system should yield the sign of the Doppler shift: i. e., whether the missile is approaching or receding.

Fig. 4 is a block diagram of a receiving and recording system currently being developed for application in the satellite ionosphere experiment. At first glance the system does not appear to differ greatly from a conventional superheterodyne receiving system. Further examination of the characteristics of the various components will, however, reveal several distinct departures from conventional techniques of receiver design.

Considering first the general design it is to be noted that the receiver is in fact a dual channel device employing triple heterodyne conversion. The received radio frequency signal, 74 mcps, is reduced successively to intermediate frequencies of

approximately 10 mcps, 260 kcps, and 5 kcps. A highly stabilized reference frequency, also at 74 mcps, is similarly treated in the reference channel of the receiver. It is to be noted that the desired intelligence, the difference between the reference and signal frequencies, has not yet been extracted or "detected."

The bandwidth of the system has also been reduced in successive stages. Since a reasonably high degree of phase fidelity is desired and phase transients in the input to the tracking filter are undesirable, the 5-kc IF amplifier will most likely be adjusted to have passband half-power points at about 1 kcps and 9 kcps giving a bandwidth of 8 kcps. Final reduction of bandwidth to the required fraction of a cycle per second should be accomplished prior to final Doppler signal detection for two reasons. First, the cross modulation products that might be introduced by the detection process will not increase the noise components in the filter input. Second, the filter will not be required to handle frequency components at or near DC which would be present after detection. Since the signal frequency may have any value between 2000 and 8000 cps ($5 \text{ kcps} \pm 3 \text{ kcps}$) and the filter passband must be centered on this frequency, the filtering device will be required to "track" the incoming signal, i. e., the center frequency of the passband must be adjusted continuously to the instantaneous value of the signal frequency.

Such filters are generally referred to as active or tracking filters. In practice, filters of this type usually consist of a local oscillator whose frequency is compared to the incoming signal frequency and is continuously adjusted to equal the signal frequency. Fig. 5 is a block diagram of such an electronically synchronized or "locked" oscillator. If the signal comparing device is a frequency

sensing type, the oscillator is said to be frequency-locked. If a phase detector is employed, a phase-locked oscillator results. By appropriate choice of the transfer function Y_p the bandwidth characteristics of the filter may be varied. It is proposed to use such a phase-locked narrow band tracking filter at the output of the 5 kcps intermediate frequency amplifier. Note that by accomplishing final bandwidth reduction at this point the full advantage of predetection filtering is realized.

It was pointed out earlier that the filter bandwidths, as determined by \dot{f} and S_v/N_v considerations, range between 1 kcps and 0.4 cps. The characteristics of the tracking filter for this application are being designed to provide for a choice of twelve bandwidths ranging from 0.1 cps to 128 cps in binary steps.

The transfer function is also being designed to provide essentially zero phase tracking error for a constant rate of change of signal frequency ($\ddot{f} = 0$). For variable rates of frequency change the phase error between the signal and the locked oscillator will vary but in no case may it exceed 90 degrees or the oscillator will fall out of lock and the filter will cease to function properly. Provisions are being made to indicate, by means of analog voltages, the first, second, and third derivatives of signal phase and a correlation coefficient proportional to the phase difference between the input and output signals of the filter. By observation of these indicating devices an operator may manually select a bandwidth appropriate for the existing conditions.

The sign of the Doppler frequency is established by determining if the instantaneous frequency is higher or lower than the reference frequency. This can most conveniently be accomplished before the reference and signal carriers are combined to yield the

Doppler frequency. The sign indicator is shown in Fig. 4 as being in parallel with the phase detector in the magnetic tape recorder output circuit. It could be used with equal effectiveness ahead of the recorder. The sign indicator is of conventional design, making use of a quadrature relationship between the reference and signal frequencies to determine phase lead or lag of one with respect to the other.

The filtered signal voltage and the reference voltage may be compared in a phase detector to derive the desired intelligence - the Doppler frequency. Since the reference voltage is essentially uncontaminated by noise, the phase comparison process yields nearly optimum cross correlation detection.

The signal frequency can also be recorded on magnetic tape or could be transmitted to a central recording station before being recorded. Since the signal information is still contained within the band of frequencies from 2.0 to 8.0 kcps a transmission system capable of transmitting DC is not required. It will of course also be necessary to record the reference frequency so that the reference and signal frequencies may later be combined in a phase detector to extract the Doppler frequency.

An auxiliary magnetic tape recording channel is used to record the signal frequency before it has been passed through the narrow band tracking filter. The signal at this point is still heavily contaminated by noise. Since the prefilter bandwidth is about 8 kcps the S_v/N_v ratio will be 0.05 or -26 db. The noise generated by magnetic tape recorders, neglecting for the moment wow and flutter, is generally about 40 db below maximum recording level. Thus, tape noise will not seriously deteriorate the stored signal information.

The purpose of recording this noisy signal is to guard against complete loss of data in the event the tracking filter should fail to lock on the signal initially or if due to any reason it should drop out of lock during passage of the satellite. The first of these events is more likely. In any event if the noisy signal and reference signals are stored on magnetic tape they may be played back at a later time, passing the noisy signal through the tracking filter as many times as may be necessary to achieve lock-on and then combining them in the phase detector to obtain the Doppler signal information. This technique not only permits one to try a variety of adjustments of the tracking filter characteristics but further, if lock-on cannot yet be achieved, due to very high rates of change of signal frequency during the early portions of the record, the record can be played back in the reverse direction so that lock-on can be achieved under conditions of best signal-to-noise and lowest signal acceleration.

It is hard to believe that one of these post-real time filtering and detection processes will not work under the conditions to be expected. However, should this occur still another technique is available to permit reaching down into the noise to extract the signal information. Sometimes referred to as "dead signal" detection, the method makes use of the fact that one can substitute time for bandwidth and by the use of autocorrelation techniques can very effectively separate signals deeply imbedded in noise. A study (Ref. 3) of such techniques has been carried out by the Stanford University Research Institute under sponsorship of the Ballistic Research Laboratories and techniques and equipment for accomplishing such dead signal detection have been proposed.

While all of the equipment required for this proposed system has not yet reached the stage where test results can be quoted

there is sufficient evidence at hand from previously completed developments in these fields to offer a high degree of certainty that very low level signal energies of the type assumed here can be satisfactorily detected. Within the past few years a similar type of receiving system was developed for application to a program involving a projectile too small to carry a source of radio frequency energy. The instrumentation system that was developed depended upon energy reflected from the projectile. In the tests that were conducted a projectile illuminating source of only 35 watts of power at a frequency of approximately 125 mcps was used.

Fig. 6 shows both the theoretical and measured signal energies as a function of ground range from launch. It is to be noted that the measured energy level is slightly higher than the theoretical value but is always within 6 db of the latter. The tracking filter, with an effective noise bandwidth of 40 cps, first locked on the signal at a level of -177 dbw and continued to track while the signal increased to a higher level and then decreased to a level of -180 dbw. Recalling that the proposed tracking filter for the ionosphere experiment will have a bandwidth of 0.4 cps or one-hundredth of 40 cps, the noise power

level should be 20 db less, enabling this filter to work down to a level of -197 dbw. Since the theoretical minimum received power is -178 dbw the proposed system should yield a S_P/N_P margin of 19 db or a S_v/N_v ratio of about 10/1 which checks with the value of 10/1 previously predicted.

Fig. 7 is a reproduction of an oscillogram showing the noisy signal at the input to the tracking filter having a bandwidth of 40 cps and the filtered output signal.

The correlation function, an analog voltage, gives a continuous indication of how well the output signal from the oscillator is being held in phase-lock with the input signal and may be used as a measure of confidence in the validity of the filtered Doppler signal output.

The experience gained through the development and field test of this prototype phase-locked tracking filter leaves little doubt that the characteristics required in an improved filter for the satellite ionosphere experiment can be achieved. The achievement will constitute an advance of the state-of-the-art and will bring us one step closer toward the attainment of electronic instrumentation systems capable of tracking very long range missiles - and space ships.

REFERENCES

1. Berning, W. W., "The Determination of Charge Densities in the Ionosphere by Radio Doppler Techniques," Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland. Published in "Rocket Exploration of the Upper Atmosphere," Pergamon Press Ltd., London.
2. Goerke, V. H., Unpublished correspondence.
3. Honey, John F., "Detection Techniques in Doppler Tracking Systems," Final Report, Contract DA-04-495-ORD-278 for Ballistic Research Laboratories, Aberdeen Proving Ground, Maryland.

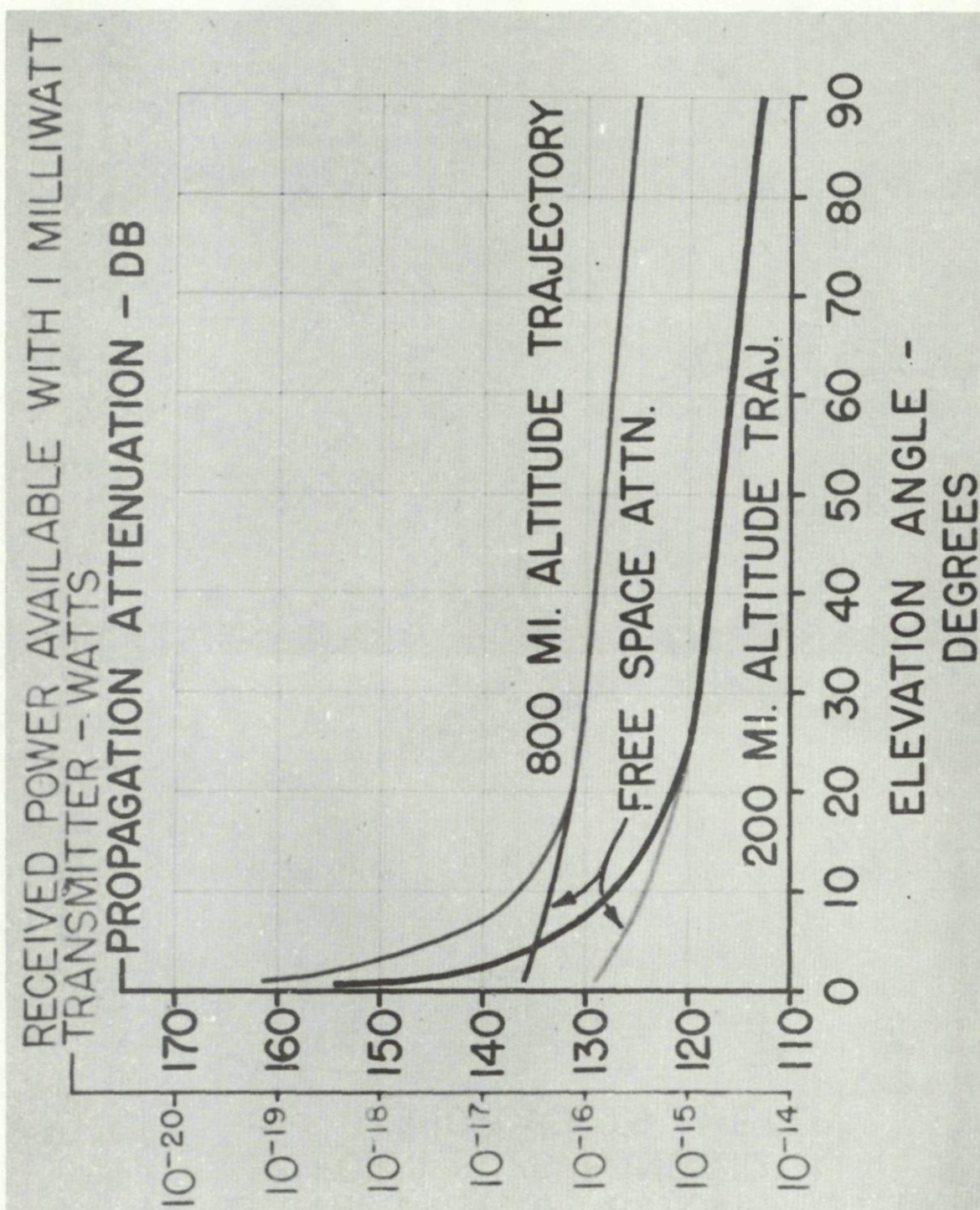


Fig. 1. Propagation attenuation vs. elevation angle.

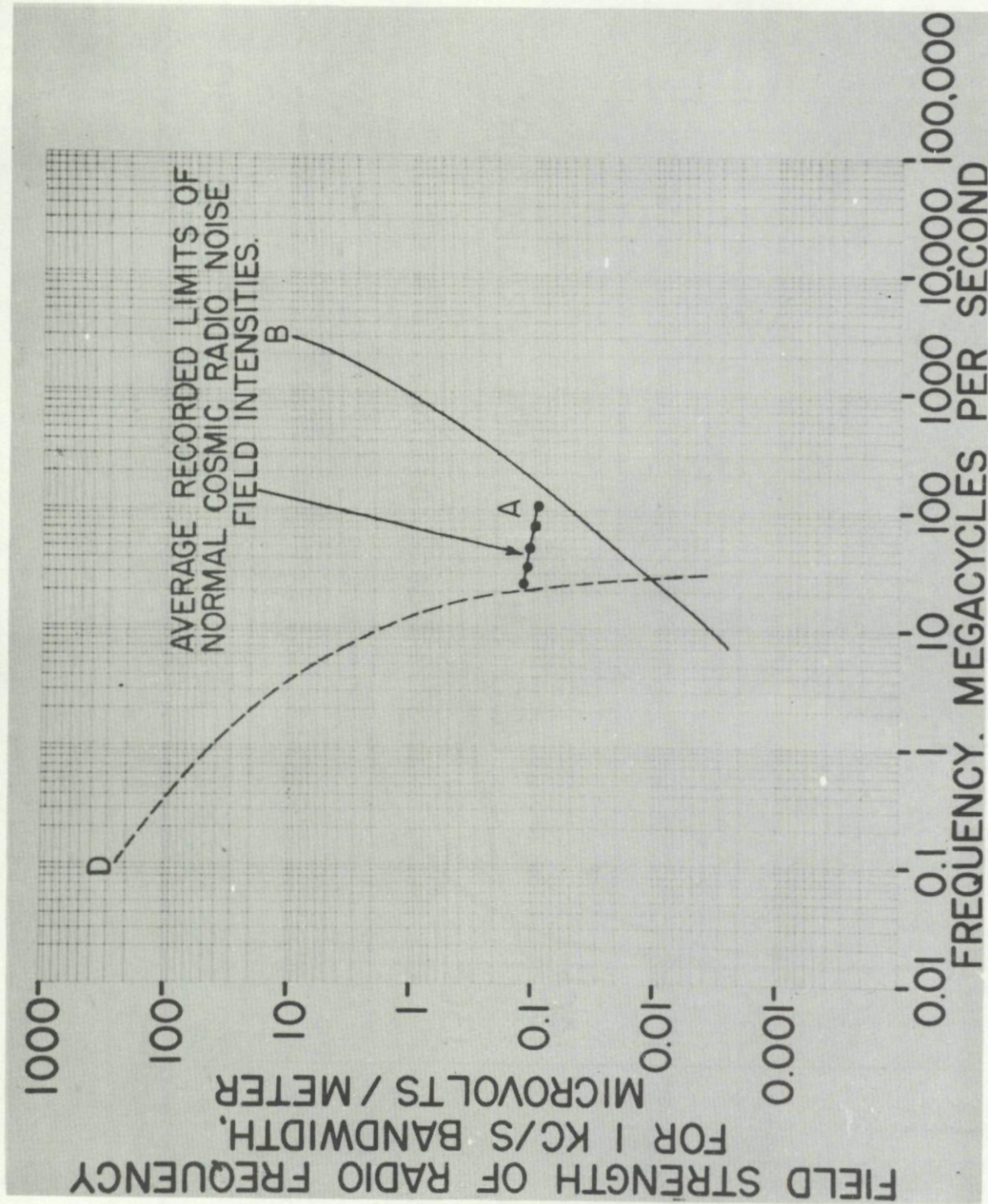


Fig. 2. Noise vs. frequency.

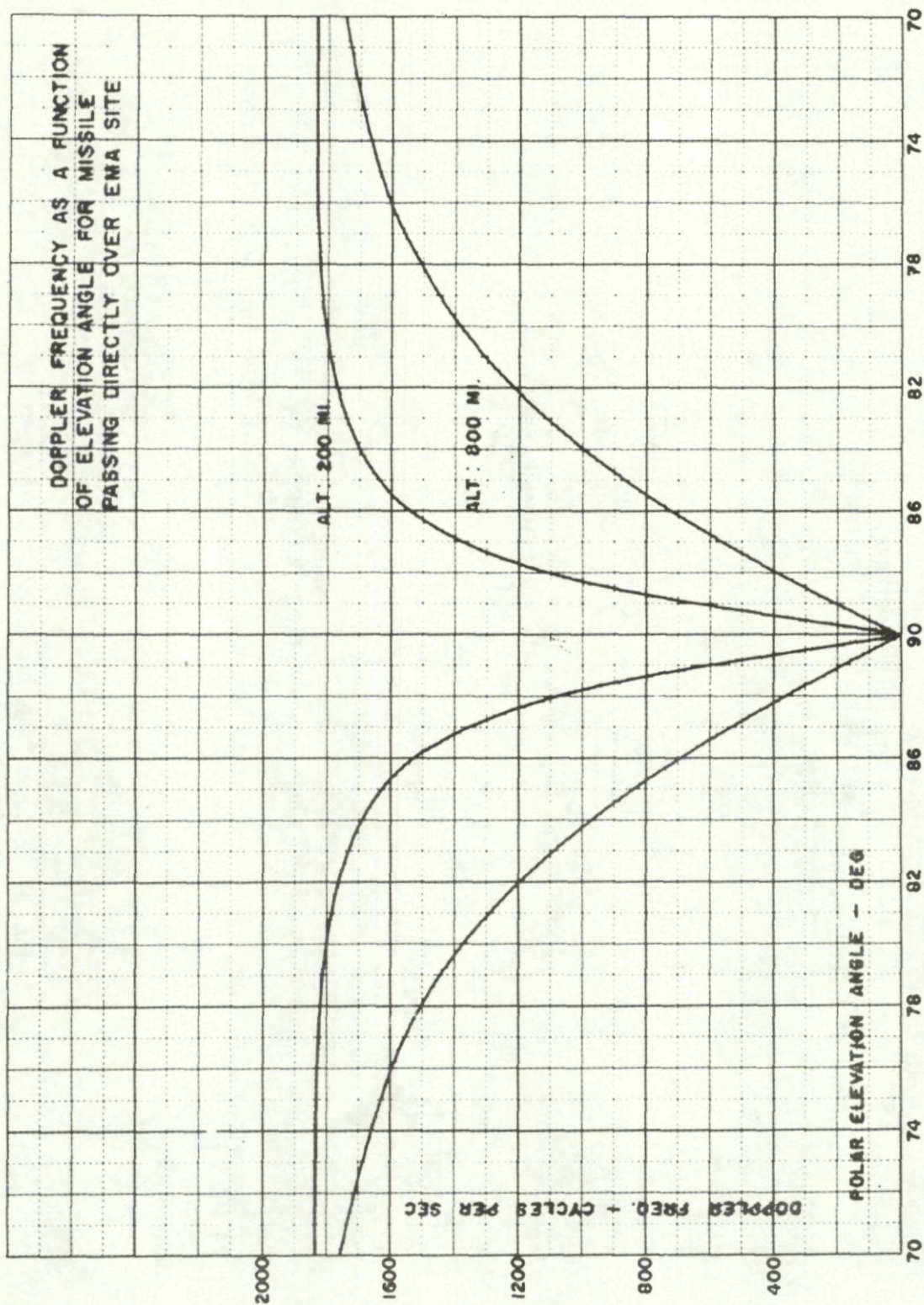


Fig. 3. Doppler frequency vs. polar elevation angle.

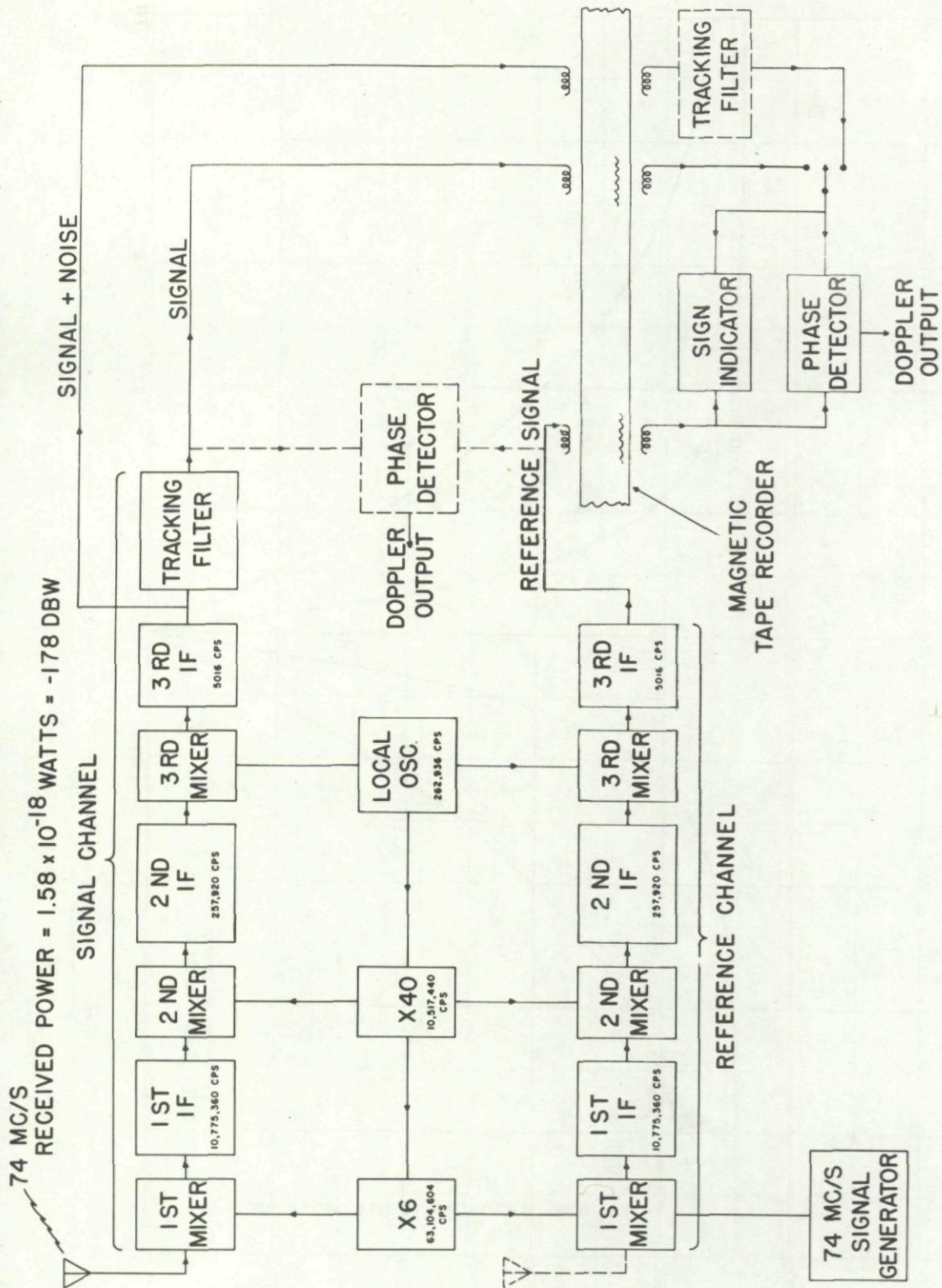


Fig. 4. Proposed satellite signal receiving system.

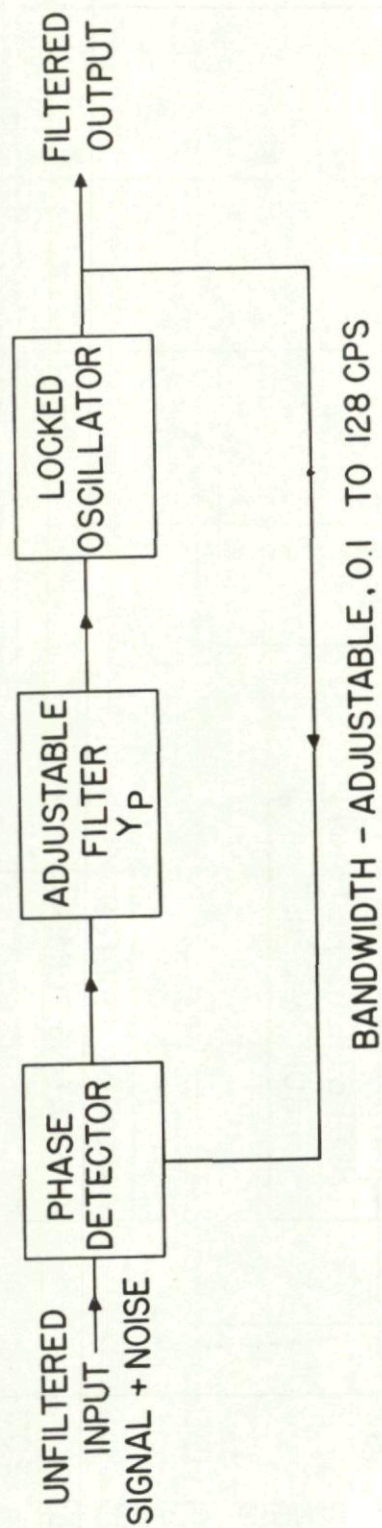


Fig. 5. Phase-locked tracking filter.

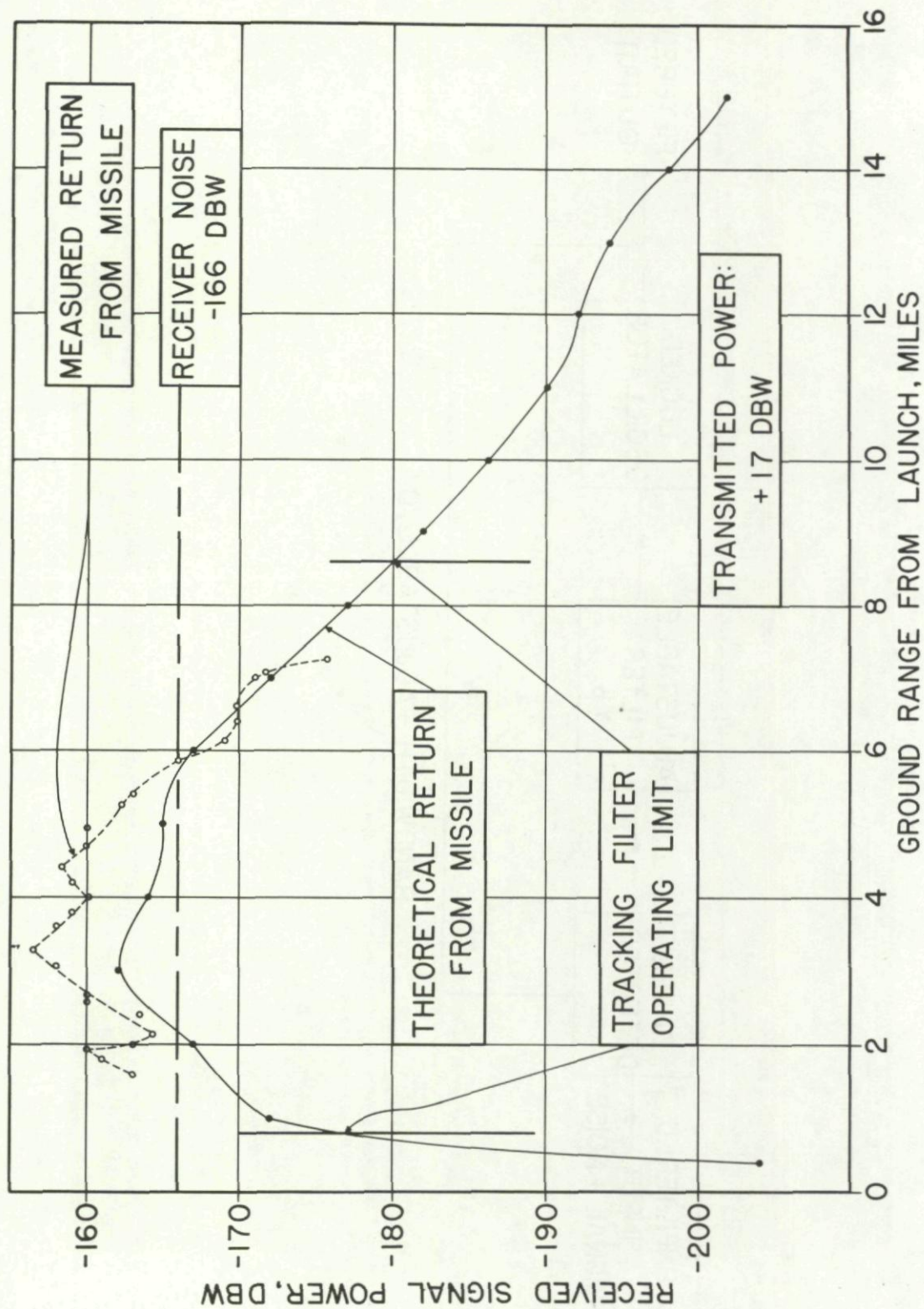


Fig. 6. Prototype tracking filter test data.

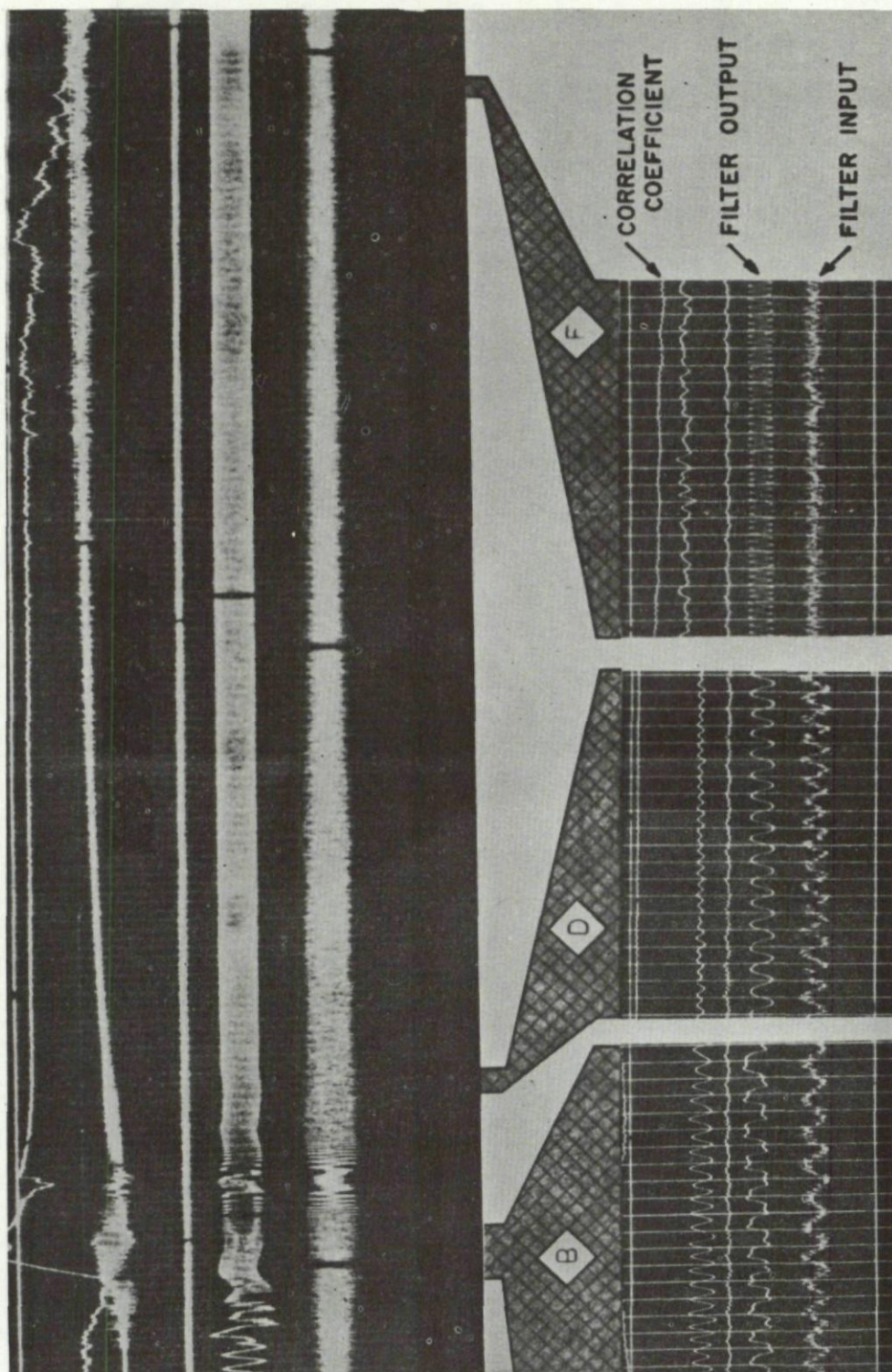


Fig. 7. Prototype filter, input and output signals.

ON THE WAY TO AUTOMATED PROCESSING OF FLIGHT MEASUREMENTS

Dr. W. E. Klemperer*

SUMMARY

This paper presents a survey of the methods available for automated measuring and processing of flight data. The measurements necessary to describe the behavior of a test missile are divided into three groups, namely, those that can be obtained from a ground station, those that can be taken only by devices located on the test missile itself, and those that can be obtained conveniently from another aircraft. The methods available for each group are then discussed separately.

SOMMAIRE

Dans cette note l'auteur donne un tableau general des méthodes utilisables pour la mesure et l'évaluation automatique de l'information en vol. Les mesures nécessaires pour decrire le comportement d'un missile en essai sont divisées en trois groupes, celles pouvant être obtenues à partir d'un poste au sol, celles pouvant être enregistrées seulement à l'aide d'organes placés dans le missile lui-même, et celles qui peuvent être aisement obtenues à partir d'un autre engin volant. Les méthodes utilisables dans chaque groupe sont ensuite traitées separement.

While the scope of this symposium is Guidance and Control and a majority of the papers presented deal with the theory of closed loop guidance and control, the subject of this paper concerns those test flight measurements from which there is no feedback to the same flying vehicle. However, there should be feedback of sophisticated intelligence from one test mission to a subsequent mission in order that a series of flight tests may reflect systematic progress. Obviously then, there must be some time allowed for the analysis and digest of the data gathered on any flight and for the application of the lessons learned. Even so, the pace of development of new projects is such as to demand that this digest time be minimized and that all those chores be mechanized

which do not require human judgment. The value of elaborate flight test instrumentation has only gradually been realized.

Lord Kelvin has said that we can begin to understand only those phenomena which we have learned to measure quantitatively. This certainly applies to the development of guided missiles. The faster our vehicles have become, and the more detached from Mother Earth, the more we must rely on instruments that are quicker and more precise than the five senses with which nature endowed us. In the earliest aircraft which flew at speeds less than 100 mph, the pilot was able to observe and jot down on a kneepad the few measurements that were necessary to determine the basic performance and

*Douglas Aircraft Company, Inc., Santa Monica, California.

control behavior of his craft. However, with new supersonic aircraft things happen too quickly, and in unmanned missiles which are guided remotely or automatically there is nobody along to read gages. Yet their greater complexity and precariousness calls for increasingly intricate and elaborate measurements to develop new systems, to prove their proper function and the reliability of all components, or to discover unforeseen difficulties and find remedies for them.

During the development of the German V-1, it was stated that nearly 500 were launched in instrumented practice and test flights until an acceptable degree of reliability against all hazards was achieved. In general it is now pretty well accepted that any new flight vehicle, manned or unmanned, must go through an intensive development stage during which elaborate instrumentation must be provided to bring home a complete record of the function of all its system components under systematically varied conditions so that the most appropriate balance of all significant parameters can be established.

The measurements which are needed to describe fully the behavior of a vehicle in flight can be divided into three groups from the viewpoint of instrumentation logistics:

- (1) Those that can be observed and measured from a ground station directly.
- (2) Those that can be taken only by sensing devices located on the flying test vehicle itself.
- (3) Those that (sometimes) can be advantageously measured from another aircraft.

The first group covers the determination of the trajectory of the vehicle with respect to the earth, by means of optical methods such as phototheodolites or cinetheodolites or by radio or radar. Ground velocities can be derived (within limited accuracy) by differentiation from position coordinate histories or directly determined by Doppler techniques.

The second group comprises all measurements of the operation of components, (control surface excursions, torques or moments, forces, temperatures, pressures, fuel flow, airspeeds, angles of attack and yaw, accelerations, and rotation and attitude) as well as ambient conditions (such as external air pressure, temperature, and other aerological properties).

The third group is particularly applicable to the guidance of anti-aircraft test missiles to an air intercept of a real flying target where proof of the end game must be based on measurements of the relative motion of the two flying objects in close detail.

Let us consider first the trajectory determination from ground stations. For artillery purposes the time-interrupted tracking of ballistic shell trajectories by means of fixed-plate cameras and photogrammetric evaluation of the pictures taken from two conjugated stations at first sufficed; it has been practiced for many decades. The greater resolution and accuracy desired for the recording of aircraft maneuvers and for anti-aircraft work led to the development of ballistic strip cameras for the takeoff phase and cinetheodolites (notably by Askania in Germany and by Mitchell in USA and later by Contraves in Switzerland) for the flight phase. With these instruments, the flying object is manually tracked by watching it through a telescope while aiming a telescopic camera approximately at it. At regular intervals, pictures are taken which furnish a

telephoto image of the object against a cross hair and also a photographic imprint of the scale readings of azimuth and elevation angles of the telescope axis.

From these pictures, the time histories of the true azimuth and elevation of the lines of sight to the object from two or more surveyed theodolite stations are determined and the true position history of the object is reconstructed at suitable intervals by triangulation routines from simultaneous sets of such angles. Redundancies are resolved by least squares or weighted averaging methods. Longhand, this was a very tedious procedure. Punchcard machines and later electronic calculating machines reduced the time and tedium of the trigonometric calculation to a very acceptable minimum. In order also to facilitate and expedite the reading of many individual cinetheodolite frames, several steps were introduced at the various test ranges, beginning about seven years ago. These innovations very materially reduced the time required to reconstruct a trajectory.

The first advance was accomplished by an instrument called the Iconolog, a projector with mechanically manipulated cross hairs and a pair of analog-to-digital converters, the x and y coordinate output of which was automatically transferred to a relay memory and thence to an electric typewriter or an automatic card punch or both. The cross hair was positioned over the image of the object by hand and eye, deliberately, to employ judgment in allowing for the different appearance under different conditions of range aspect, illumination, and contrast and to insure tracking the desired part of separate stages. The machines were also equipped with keyboards on which the reader could key in all numerical information read off the picture such as azimuth, elevation, attitude, run number, and picture number (or time). This information was retained

on a short memory and then entered automatically on the typewriter record and/or on the punched cards. In later machines the reading of azimuth and elevation scales was reduced to manual scale matching which relieved the operator of the task of interpolating and keying numbers.

Machines of this general type were developed by Douglas, Genisco, Benson-Lehner, Coleman, Telecomputing Inc., etc., and the same idea has been applied to microscopic comparators for the reading of the films taken with ballistic cameras. The machines differ in various mechanical and optical features. Most of them display the magnified picture on a translucent or opaque screen (which is less tiring than looking through a microscope); some move the film on a microscope stage, others move a cross hair behind the screen; and still others project a luminous index spot upon the picture. Some manipulate this motion by hand; others motorize it at variable speed. Where it is desired to record the apparent attitude of the object in the picture, an angle-measuring feature is added, either in the form of a turntable which rotates the film projector or the film gate, or a cross hair or luminous line. This rotation is then digitized, committed to a relay memory, and recorded automatically.

Of course, it is tempting to do away with the human link in the reading of the film. This is readily feasible only when the object appears as a well-defined point of contrast against its background. Some type photoelectric sensing device and a television-like scanning procedure can be employed, but the proper adjustment to the dimensions of the spot and its contrast is a delicate problem at best and requires some human supervision or monitoring. It is even more tempting to do away with the tracking error altogether, and thus eliminate the necessity of reading

or gaging the apparent displacement of the object image from the optical axis of the tracking theodolite so that the whole photographic step can be dispensed with.

The introduction of velocity feedback for "aided tracking" went a long way to assist the trackers in following their objective through the sky. Although some designers of equipment feel confident that perfect manual tracking can thus be accomplished, others rely on wholly automatic tracking which positively eliminates human errors and fatigue. Just as automatic star trackers have been made successfully, so a flare carried on a missile can be tracked automatically and accurately by means of a photoelectric seeker at night. It is only a matter of providing a good balance between resolution, slewing speed capability, and damping.

In daylight, it is more difficult to supply sufficient light intensity and contrast especially for high altitude or long range work. Infrared seekers can also be built to lock on missiles carrying a hot flare. However, once the target photography is rendered unnecessary, it becomes desirable to eliminate the photographic process entirely and substitute other means for the retention of the theodolite angle information, to digitize it, and automatically record it at suitable intervals in real time on the fly. This is an ambitious enterprise because the resolution required is very high; to obtain resolution of 1/10 mil would require 64,000 digitized stations along one complete circle, and even this furnishes an accuracy of no better than about 10 to 50 feet at 20 to 100 miles range.

The ultimate in tracking devices should be the modern antiaircraft radar. It furnishes its own range information and requires no conjugation of several stations; but to achieve angular accuracies exceeding those of optical telescopes it requires rather

elaborate installations on the ground, and, of course, transponders carried on the missile. The modern all-electronic automatic tracking radar furnishes its own graphic record of the computed flight trajectory of the tracked object immediately, and it can be equipped to plot velocity components as well. Where a higher accuracy or resolution than that given by a large automatic plotting board is required, the tracking angle and range information is transformed into digital form and recorded as such at suitable intervals.

However, it often pays to equip even a fully automatic tracking radar with telephotographic backup instrumentation. A picture, which according to a Chinese proverb is worth 10,000 words, and according to the television engineer about 300,000 points, can tell the story of a dramatic event such as the separation of a missile from a booster stage, or an accidental breakup more vividly than a coded signal. The pictures can also serve to verify the true tracking even if they need not be photogrammetrically evaluated. Otherwise, they can be evaluated as "bore-sight" pictures in the same manner as cine-theodolite pictures to tell the story of the tracking error and to furnish clues for improvements of the tracking system. At any rate, it can be anticipated that within a few years, fully automatic ground-to-air trajectory and velocity measurement equipment will be available at all major test ranges for aircraft and missiles; eventually it will become standardized. In the meantime, much work will continue to be done with present-day semiautomatic equipment in which the human observer, aided by devices that reduce boredom and fatigue, supplies judgment and intelligence.

Radial components ground-to-air velocity measurements by Doppler or other signal phase comparison techniques such as Raydist

can be recorded in true time. The records of phase difference measurements, however, which must be referred to hyperboloidal coordinates in order to reconstruct trajectories, require counting and considerable processing and computing. One must exercise careful judgment and/or take advantage of redundancies whenever any record is interrupted. Electronic calculating machines are indispensable here; semiautomatic trace reading aids are in use but complete automatization will depend on perfect reliability.

Let us now turn to the second category of measurements; namely those which must be gathered by instruments carried aboard the test vehicle. Such instrumentation must be designed to be accommodated on the vehicle without interference with its function. Hence, airborne instrumentation cannot simply be bought and put aboard; it must be planned for in the early design stage. As a matter of fact, during its brief history, the philosophy of airborne flight test instrumentation has undergone drastic changes. In the beginning, people were reluctant to load airplanes or missiles with instruments.

When the necessity was recognized, test engineers were at first satisfied with putting relatively primitive and inexpensive instruments aboard considering that the investment would be of uncertain value at best; little or no attention was paid to the cost in money, time, and effort which later would be spent to evaluate the results. Gradually, the problems of making and installing reliable transducers were solved and efficient methods for the transmission, reception, and preservation of records were developed. This soon led to the accumulation of great masses of data which eventually grew to such proportions that it became impossible to keep up with them, to read all of them, to organize and plot the results, and to interpret and evaluate

them in time to serve a useful purpose before engineering decisions regarding adjustments or changes for the next tests had to be made.

The magnitude of the job is readily realized when it is considered that in one now standardized method of telemetry about 1000 data are collected per second so that in a single flight of two to ten minutes, 100,000 to 500,000 measurements from this source alone are collected. They would have to be read, sifted, decommutated, processed according to calibration, plotted, indexed, cross plotted, and correlated with information gathered by other instrumentation in computations of response characteristics and performance to exploit fully the wealth of information garnered. A whole new industry has developed as the result of various efforts to facilitate these tasks and to overcome delays by reducing the amount of human effort involved in these processes through progressive mechanization and automation. This development and its signal success now begins to exert its own influence upon the planning of instrumentation for new projects and new facilities. Obviously, it now pays to so design the instrumentation in the first place that it lends itself better to expeditious automatic evaluation even if the installation in the individual vehicle costs more.

The earliest technique perfected for obtaining frequent records of instrument indications aboard a flying vehicle was to take motion pictures of indicators arrayed on an instrument panel and suitably illuminated. This method, still standard in many places, was developed to routine perfection for manned aircraft. From there it was adapted to unmanned missiles, but here the problem of salvage of the film records arose where no provision for an intact landing could be made. Films were protected by armored cases and retrieved from the impact wreckage, in some instances by parachute recovery, but, in any case, a tedious and cumbersome process. Once the films were

delivered and processed, they had to be read, when magnified by a projector such as a Recordak, on a screen conveniently positioned in front of the reader.

Here, too, mechanical aids such as a luminous pointer manipulated into coincidence with the screen image of any dial or scale instrument and then automatically digitizing and recording the reading (as in the Benson-Lehner Boscar) on the basis of preset zero and scale calibration helped significantly to expedite the reading process and reduce human errors; it was an important step toward automatization although still quite short of complete automation.

Instead of photographic registration, the instrument indications were sometimes converted on board into frequency modulated or digitized signals and recorded on magnetic tape. The retrieving chore remained, but the danger of losing the record by accidental exposure of the film was avoided and the time required for developing film was saved. A radioactive source carried on the recorder case was found useful to facilitate locating the impact place on a vast test range. Once magnetic tape was salvaged, the record on it could be played back, processed, and displayed on all-automatic devices without further human effort. This method is still standard with some experimental missiles.

The retrieving problem was eliminated by radio-telemetry which, during the last 10 years, has gradually replaced airborne recording. Several systems have emerged into practical service and attained acceptable reliability. They are usually distinguished by the method of multiplexing the many individual measuring channels in the signal transmission over the radio carrier link. These methods lend themselves differently to mechanization or automatization of the

data reduction which comprises the steps of unscrambling (demultiplexing), rearranging, linearizing, reading, plotting, and utilizing the data.

The most widely accepted multiplexing methods are (a) sharing a frequency band by subcarrier frequency modulation, (b) time sharing by commutation (successive sampling), and (c) by combinations of these. The choice of the multiplexing method also affects the choice of the types of transducers to be installed aboard the vehicle. Transducers are those instruments which translate the original response of the measuring sensor (pressure syphon, movement gage, torque-meter, dynamometer, Bourdon tube, thermometer, dilatometer, flow gage, vane, accelerometer, gyro, etc.) into an electrical quantity such as a DC voltage; hence, a condenser charge, or an inductance or reactance, and hence an electric oscillation frequency.

From another viewpoint, the various measurements may be distinguished as to the resolution in quantity and frequency required. Some measured values change but slowly and monotonically; e.g., the supply of fuel, pressure medium, etc., so that infrequent sampling suffices. Others may vary fairly rapidly in controlled maneuvers or in the performance of automatic stabilization or remote guidance; e.g., control movements, torques, forces, accelerations, angles of attack; and, to a lesser degree, airspeeds, ambient air density, and temperature. Finally, there are those that attain high frequencies from 30 to over 1000 cps; viz., flutter, vibrations, and structural oscillations. These are better caught by continuous recording than by sampling.

For continuous recording, it has been found practical to multiplex six and up to 12 instrument channels over one carrier link, by the FM-FM technique. Each channel is assigned a sonic subcarrier band; the centers

of the bands are usually staggered by a factor of 1.5, more or less, depending on the balance of number of channels required and the resolution desired of each. To good advantage, the channels requiring the highest frequency response are assigned the upper frequency bands. For safeguard against cross-talk, the sum of all channel responses available is less than the transmission link bandwidth. The telemetry signals are received on the ground and either stored on magnetic tape or immediately converted to true time graphs on strip recorders or both. The subcarrier components are separated by suitable filter circuits and individually and continuously recorded so as to furnish an immediate qualitative graphic picture of the history of the flight measurements.

These graphs may still be inconvenient in scale, nonlinearity, zero points, crisscrossing, and length. They may be anamorphically reproduced to other time and ordinate scales in automatic reprinting machines (such as the Genisco-Panoramagraph). Above all, the original records may be manually corrected at as many points as desired, by transposing them, linearizing them, adjusting them for zero point and scale value for in-flight calibration (by machine-aided manual operations, for instance on the Benson-Lehner "Oscar") and by replotting them automatically to the desired scales by means of a digitized output memory and an automatic data plotter.

Several years ago, it seemed that FM-FM telemetry would become standardized to the exclusion of all other systems, largely because of the ease with which it permitted graphical representation of results, despite the serious limitations in the number of instrument channels that could be multiplexed on any one radio link channel. However, this picture changed with the development of automatized data readers for reading, analyzing, and graphic plotting.

The standardized PDM system, as developed by GE and others, employed a commutator which successively switched 28 instruments on and off (plus two pauses for cycle identification) in any one revolution, while turning at about 30 revolutions per second, so that almost 900 measurements would be transmitted per second, each taking about 1 millisecond to sample. In some aircraft installations, mechanical commutation was stepped up to 88 channels and 1500 readings per second. (With electronic commutation, several thousand readings per second can be obtained, but there are limitations to the resolution attainable over available radio links since the pulse duration is inevitably shortened with higher repetition rates.)

The measured value sensed by a transducer is usually furnished as a DC potentiometer voltage. This voltage, when matched by a gradually charged condenser, determines the duration of the charging process and with it, the duration of the emission of a modulated signal over a telemetry transmitter. These pulses of various lengths are displayed at the telemetry receiver station by a sweep oscillograph and photographed on a running film. In the "Hermograph" type machine (developed about 1948) the films are projected through an optical system and automatically scanned by a sweeping mirror, past a photoelectric cell which, when it senses the end of a line, stops the charging of a condenser and thus reconstructs the analog voltage of the transducer. A patch board permits selection of up to six of the 28 channels to be read in any sequence of sorting. Each measured value is plotted automatically on a pen and strip recorder, different point symbols distinguishing different channels. Film advance and time counting are also automatic.

A bank of potentiometers is available to permit adjustment of zero, sign, and scale

value of each trace to conform to suitable units on the graph readily legible in terms of pounds, pounds per square inch, feet per second squared, degrees, degrees per second, etc., to accuracies of the order of 1/2% from input, or 1 to 3% from transducer. These machines, plotting at the rate of about one point per second, and also furnishing punched cards for later computation, were a vast improvement over previous methods involving human operators and they drastically cut down the backlog of records which remained unread before. Yet, as the development pace of missiles quickened, this speed no longer sufficed and the desire to process the records at more nearly real time grew. This became possible through the substitution of magnetic tape for photographic film as the storage medium.

Magnetic tape was first introduced as a backup device; now the roles are reversed. In Douglas' latest system, (one using a converter built by Magnavox Research Laboratories and another built by Douglas), the main records are received on tape and only a spare record is filmed as a safety measure and for visual monitoring. Presently, the magnetic tapes are collected at the range stations and mailed or flown to the central data processing facility at the manufacturer's plant, where they are played back, sorted, edited, and analyzed by engineers in close touch with the design staff. Future plans, especially in the airplane test flight division, envisage direct retransmission of the signals as received at the range via a 100-mile chain of microwave link stations to the central facility at the factory. Here the signals will be redisplayed and watched as they come in. This will be another step toward complete automation, but it is not yet in the service stage.

Once the record is available in magnetic tape form as a string of signals of varying length, concurrent with some standardized

time record, it remains to separate the individual channels by a decommutator which repeats only the signals belonging to any one selected channel and re-records it with proper time reference. This is accomplished fully automatically in a synchronized decommutator while playing the original record onto a new tape for maintenance of accuracy. Currently available magnetic tape can store up to 14 tracks per inch width. Up to about three-quarters of a million decimal digits can be stored on tape occupying one cubic inch when rolled up.

In one system in use at Douglas, the information goes directly into a digital tabular record. Otherwise the information must be converted from analog form (pulse duration) to digital form at some stage or other. Many schemes are in vogue for this purpose and many electromechanical, magnetic, and electronic devices have been developed and applied to implement them. The former are adequate for relatively slow processes or where a read-out is required only while the system is being sampled while at rest. Such converters are manufactured by Genisco, Giannini, Coleman, and Telecomputing.

Where high speeds and read-out on the fly are required, electronic converters are used. Such instruments (quoting 8,000 to 100,000 code translations per second) are being made by Radiation, Inc., Vicdar, Consolidated Electrodynamics, J. B. Rea, and Epsco. Digitizing is usually accomplished by comparing the unknown voltage with a binary series of known precision voltages, the value of each voltage being one-half the value of its predecessor. The binary digital code representing the unknown voltage is formed by assigning a binary value of one to those reference voltages whose sum most nearly approximates the unknown, and zero to all others. The residual error will be no larger than the smallest reference voltage.

PDM type data are simpler to digitize and to present than FM-FM data, because they are already received in discrete samples. One Douglas system for digitizing PDM data employs a simple fixed-frequency oscillator that is gated with the leading edge of the data pulse and closed at the trailing edge. The cycle count is then read directly into registers and the corresponding binary number is formed. The other system converts the PDM signal back to analog voltage, applies scale factors, then converts digit voltage to digits.

Once digitized, the data may be recorded on magnetic tape for use with a computer or they may be sent to a recorder containing, say, 128 fixed styli, each stylus representing one of the 128 possible quantizing levels. The rapidity with which the styli are activated suffices to present a practically continuous trace.

Analog to digital conversion is still in the flux of development. No digital code has yet emerged as undisputed standard; some translate in decimal numbers, others in binary, again others in binary coded decimal digits; the choice depends largely on the language of the computer to be used when the data are further processed.

Guided missiles are usually meant to be guided to a target. The target may be on the ground, on the water, or in the air. The success of the missile system is usually reflected by the miss vector. It is therefore of vital interest to measure this miss vector. If this is done by radar or cinetheodolite from the ground, the result is a small difference between two large vectors. It is therefore often worthwhile to measure the miss vector aloft by direct communication between missile and target during the end game with higher potential accuracy. In anti-aircraft work, the only really convincing

proof of success of the system is obtained by firing test missiles at real flying aircraft. For reasons of personnel safety, these are either tow-targets or remotely controlled drones.

It has been found expedient to equip these drones with high-speed cameras which record the near passage of a test missile and a pyrotechnical token detonation. These cameras have wide angle optics, lenses that encompass 140 to 160 degrees. Four of them in a tetrahedral array cover the entire spherical field and photograph anything that goes by. The cameras run at film speeds between 200 and 500 frames per second. They do not run in enforced synchronism, but they photograph on every frame an array of small neon lights which indicates common time in an unambiguous binary code to one millisecond resolution. Clusters of four cameras are often mounted in streamlined pods on each of the wing tips and on some large drones additional cameras are installed in the nose and tail of the fuselage. The cameras are started automatically by radio signal. Parachutes bring the camera pods to earth in case the drone is hit and destroyed in the air.

Evaluation of the pictures which show the missile is done by a semiautomatic routine. The angular coordinates of the target image are matched by the cross hairs of an Iconolog, Boscar, or similar machine. Here human judgment is needed because the missile image may change size and aspect rapidly from frame to frame. The angular position of the line of sight from the camera to the missile is then automatically punched on a card together with the time code. The rest of the calculation is then accomplished on an electronic computer in a routine procedure.

The computation comprises the correction for the optical distortion of the wide angle lens, the trigonometrical calculations to refer

from camera to drone coordinates, the decoding of time, the interpolation to simultaneous time instants between two cameras at opposite ends of a baseline such as the wing span or fuselage, the triangulation from two thus conjugated camera stations, and the plotting of the miss vector components in target coordinates. The Douglas airborne recording system has become standardized under the name Kinescore.

With air-launched missiles, it may sometimes be desirable to take motion pictures of the missile from the mother aircraft and to photograph on the same pictures the instruments which give a record of the guidance system performance. Special cameras for this purpose are under development.

While airborne photography is quite effective, it is admittedly expensive and cumbersome. Therefore, effort is being directed upon the development of all-electronic missimeters. In these, signals are exchanged between the target and the missile in radar fashion and they are received and evaluated either on a ground station or in the air, be it on the target, or on the mother aircraft. Several such systems are presently under development. They will probably eventually come into general use, but since they cannot give as many details as photography, the latter will probably remain desirable as a backup.

Because of the great variety of conditions under which missile-target encounters may occur (different target size, speed, altitude, evasive maneuvers, and at different ranges), it would be very expensive to cover all of them by drone intercept tests. Economy dictates that actual drone tests be limited to a reasonable number of typical or critical conditions, and that the rest be covered by firing tests with synthetic targets presented by a simulator with a greater or lesser degree of sophistication.

The big established flight test stations and firing ranges are busily planning and progressing toward a future when all measurements pertaining to the flight of experimental aircraft and missiles can be gathered, recorded, sorted, presented, analyzed, and interpreted, fully automatically, in a central facility, so quickly that information valuable for the next mission will be available in a matter of hours or minutes instead of days or weeks. Some plans go so far as to combine in this facility the processing of the vast output of a variety of field experiments, such as static power plant test stands and rocket sled tests, the better to utilize the inevitably expensive and elaborate data processing and computing facility. However, such long range developments should not be pressed forward too hastily. They must be projected with considerable foresight and must remain versatile and adaptable so that they can take advantage of new measurement techniques that may become practical in the meantime.

It must also be remembered that full automation does not mean complete dispensation of human judgment. Some intelligent monitoring and management will remain necessary. For instance, it would be wasteful of communication bandwidth if all available channels were equipped to handle the highest occurring frequency response and assigned blindly to channels regardless of their characteristics. An unnecessary amount of trivial data would be collected and handled for no one to examine. It is conceivable to design instruments so that the time scale is made variable and automatically opened up whenever an important event or maneuver is expected or programmed, or when an unforeseen departure from the anticipated behavior occurs. However, even this refinement will not take the place of the human intelligence necessary in planning or modifying the evaluation procedure. Therefore, so-called

"quick-look" graphical presentations are inserted at suitable steps of the procedure to enable an observer to separate the significant portions of the record from the trivial and the high priority critical material from the lower.

The "quick-look" graphical presentation is usually considered adequate as qualitative data with 5 to 10 percent accuracy. However, it is often possible to achieve much higher accuracy, in which case the quick-look presentation can be accepted as final or quantitative data. Presently, real-time or "on-line" reduction is being accomplished by taking the output of a receiver and feeding the input directly to the data system for graphical presentation. This is the first step toward full automation; the next phase would be direct input into large scale computers. The complete implementation of the trend toward automation may still take years and it is possible that it will be followed by a regression to allow more human judgment again to enter the procedures.

In a paper presented before the convention of the Institute of Radio Engineers in March 1956 and published in the IRE Proceedings, D. G. Mazur of the Naval Research Laboratory summarized the preparations being made for telemetry from the earth satellites to be launched under the auspices of the International Geophysical Year. The following is quoted from his paper: "A program embracing the use of time division and frequency division multiplexing has been planned. The equipment includes pulse position modulation, frequency modulation and pulse width modulation, commutation devices, all proven in active service in other rocket programs, under fire, so to speak. Multiple ground stations will be employed to afford backup protection in case of any ground equipment malfunction. Permanent recording of data

will be made in some instances on photographic film and in others on magnetic tape. Real-time presentation on meters and pen recorders will be available at some locations. It is expected that a vast amount of data will be collected from each rocket. Some data will be of the form that mere visual inspection will suffice. Other data will require normalization and will be reduced semi-automatically. Use of automatic data reduction methods will be made insofar as it is practical."

The plans for the tracking of the satellites were described by J. T. Mengel, also of the Naval Research Laboratory, at the same meeting and in the same Proceedings of the IRE. He said: "The Minitrack System of radio angle tracking developed by the Naval Research Laboratory utilizes an oscillator of minimum size and weight within the satellite to illuminate pairs of antennas at a ground station which measures the angular positions of the satellite using phase-comparison techniques, independent of weather, visibility, and time of day . . . Primary data presentation will be by three direct writing records of the analogy voltage of the ambiguity resolving antenna phases, and two direct writing records in digital form of the fine antenna phases, all as a function of time . . . Six pieces of data, the times of crossing the -4° , 0° , and $+4^\circ$ zenith angles, and the north-south angles that were measured at these times, will be sent to a central computing facility within 20 minutes of a tracking event, to be used in determining the orbit of the satellite . . . The central computing facility will receive data from as many as nine of these ground stations . . . Complete ephemerides will provide tracking angle and time information for the principal locations at which optical tracking stations are located, to permit the tracking equipment to acquire the satellite, as well as for most of the major cities from

which the satellite could be visible. It is not too unrealistic to predict that during a satellite event the evening newspapers will publish on their front pages three boxes, one for the baseball scores, one for the horse race results, and one for the evening time and angles at which the satellite can be picked up!"

One may well ask how European enterprise, talent, and ingenuity fit into this picture. The answers can be sought on two levels. First, it may be wise to proceed cautiously and avoid laying out extremely

elaborate facilities unless there is really a demand for high testing rate and high evaluation speed. Otherwise semi-automatic devices and optical-mechanical and electrical aids may suffice. One may actually intercept the American development on the rebound. Second, despite the technical progress made by a growing new industry there is still ample room for the improvement of components, such as transducers, commutators (especially brushless ones), antennas, decommutators, filters, playback synchronizers, high-speed digitizers, plotters, and recorders. This offers a challenge of considerable magnitude.

PAPER ON THE GUIDANCE AND CONTROL OF MISSILES

Stephen Waldron*

SUMMARY

This paper concerns the reliability of electronic equipment, particularly as it effects the reliability of guided missiles. Reliability is here defined as the probability of survival for a specified time. The stresses applied to a guided missile at the various stages of its history and which affect reliability are discussed and the conclusion is reached that a real gain in missile reliability requires investigation of failures which occur early in the missile flight.

SOMMAIRE

Dans cet article il s'agit de la sécurité de fonctionnement de l'appareil électronique, particulièrement dans la mesure où elle influence la sécurité de fonctionnement des projectiles guidés. Notre définition de "sécurité de fonctionnement" est la probabilité de survie durant un temps spécifié. Nous discutons les tensions qui agissent sur un projectile guidé aux diverses étapes de son vol, et qui influencent la sécurité de leur fonctionnement. Nous arrivons à la conclusion qu'une amélioration réelle de la sécurité d'un projectile exige une étude des faillites qui surviennent au début du vol du projectile.

This paper concerns the reliability of electronic equipment, particularly the reliability of guided missiles.

It is necessary to define "reliability." Let us accept a definition often used in military circles. A device must operate as planned for a certain time in order to be useful. Reliability of the device is defined as the probability that it survives for the needed time.

In order to measure reliability one must be able to detect failure when it occurs. Assume for the moment that one can identify failures as they arise.

Failures in complex equipment often occur in such a way that their causes are difficult

to identify. As a consequence, failures are often classified with relation to time or density of failures per unit time, rather than with relation to cause. Failures are usually divided into the three categories: initial, chance, and wearout failures.

Early operation nearly always exposes component defects which cause the equipment to fail. Such initial defects are built in or are developed by the rigors of transportation or rework at the consumption point. After fairly fixed operating times the equipment may fail due to wearout failures in such components as batteries and rotating machinery. The equipment may fail, however, as the result of component failures which are due neither to infant mortality nor old age. For many components the probability

*Office of the Chief of Naval Operations, Washington, D. C.

of failure is independent of the cumulative time the component has operated; failure is a chance event. This is typical of vacuum tubes, for instance, which may "go" at any time.

When equipment failure results from chance failure of components, the equipment reliability can be related mathematically to the reliabilities of the separate components. The mathematical relationship expresses the fact that a device composed of many long-lived components will itself have a relatively short mean life. This is not surprising. In fact, the equipment mean life is dominated by the weakest components.

Much of the effort to improve reliability of electronic equipment has gone into determining and increasing the mean lives of components. This needed effort has generated vast failure-reporting systems and data-handling facilities. As a result, attention has been focused on those components which most need improvement.

Now consider the basic question, why do components fail. A component fails because it is subjected to stress, through a period of time. In general, failures result sooner from large stresses than from small ones. For a fixed stress and time, the probability of failure will have a distribution of values for a collection of supposedly identical components. Stated differently, stress resistance varies among "identical" components in a collection, due to differences in materials, assembly, and other factors. By combining the effects of random variations in stress resistance and random variations in applied stress, it is possible to account for chance failures of complex equipment.

The theory of chance failures outlined above applies satisfactorily to most electronic equipment. Equipment stresses are small, fairly uniform, and well enough known

so that components may be tested appropriately during design and manufacture. As a consequence there are no spectacular differences in the mean lives of components involved, and equipment failures occur randomly.

The theory of chance failures must be augmented by other considerations when guided missile reliability is considered. Guided missiles differ from more conventional electronic devices both in complexity and in the environment which they experience. The guided missile must not only handle information, but also control its own energy resources. The circuitry is often working near the limits of its capabilities and there are strong interactions among components. In addition, the guided missile environment is much more severe than that in which conventional electronics operates.

Consider the stresses applied to a guided missile at various stages of its history. Table 1 lists these stresses for a typical rocket-launched missile.

Under mechanical stress, shock (or impulse) could result from rocket launch, dropping, catapulting, or arrested landing. The other mechanical stressing should be self-explanatory. Electrical stress includes the effects of transient overloads of voltage or current. Thermal stress includes high ambient temperature and, perhaps, thermal gradients. Chemical stress is meant to cover the corrosive and other effects of chemicals in the atmosphere. Pressure, reduced at high altitudes, is effective in allowing corona and making sealed units breathe in chemicals. The stress of man will be brought out shortly.

Let us consider possible causes and remedies for failures which occur at various points in a missile's life. Following production, a missile is transported, received, and

Table 1. Stresses Applied to a Missile During its Consumer Life

Stresses		Transport	Storage	Test	Repair	Use
Mechanical	Shock*	x			x	x
	Vibration*	x	x			x
	Normal Function**			x		x
Other than Mechanical	Electrical			x		x
	Thermal			x		x
	Pressure					x
	Chemical***	x	x			x
	Man			x	x	?

*Applied or flight-induced.

**Includes hydraulic systems as well as aerodynamic controls.

***Carried by the atmosphere.

inspected. The consumer examines and tests the missile and repairs its "initial" defects. The latter are often such mechanical flaws as cold-soldered joints which have been vibrated loose in transport. A few initial defects are found in electronic components which may have had their electrical or thermal stress limits reduced by transportation and handling. The cure for such troubles may lie in better motivation of the fabricators and inspectors, coupled with better mechanical design of the missile and its shipping container.

The consumer now stores the missile. He hopes that its storage life will be long. In order to ascertain that the missile is still operable after storage, he tests it. Let us suppose that there is low confidence in the missile's ability to last through storage.

Such is often the case. Given this low confidence, the consumer tests his missile after only a few days. At this time the missile often fails to check out properly. It is therefore concluded that the missile is unable to withstand the rigors of storage. It can be seen what this means in terms of man-hours spent testing and repairing missiles.

The misinterpretation of storage effects comes about because we have here another example of the uncertainty principle. A "good" missile is defined operationally. A missile cannot really be defined as "good" in storage, because it must function to be good. When it functions, the test procedure may disable it. The consequent failure must be examined carefully and should not be ascribed to storage unless such a failure has never been known to result from testing.

It has just been implied that a missile remains operable while left in storage. This is true for long periods of time, provided that atmospheric-borne water and other chemicals are kept to low levels. Given proper atmosphere, it is found that storage life of a missile grows as rapidly as one has the courage to lengthen it.

Let us examine the test process further. When characteristic electromechanical and thermal stresses are applied, chance failures reduce the mean life of a typical missile by a factor of two or three, compared to a more conventional device made of the same components. This implies that the missile test environment is one of increased stresses, which is no great surprise.

The stress of man is brought to bear during the test procedure. First, man often prods or gouges in the wrong places while adjusting the missile or preparing it for test. Worse, he may incorrectly interpret the test results and start an unneeded repair effort. The latter, involving disassembly and handling, often produces defects where none existed before.

The problem of false diagnosis is of course two-pronged. However, perhaps out of respect for missile costs, the test man will more often repair a good missile than pass a dud. The latter effect has been found in studies of vacuum tube failures.

The problem of misdiagnosis of missile tests seems insoluble. With the present complicated test equipment, the only way to reduce possibility of error is to take more time. More time brings the inevitable true defect closer. A missile which needs no testing seems an unrealizable goal.

The dangers of creating new defects while repairing old ones have been mentioned.

One of the difficulties of the chance-failure approach to missile reliability may now be apparent. That approach has led to the collection of myriad data on failures during missile life before use. Unfortunately, the records of true chance failures are diluted by a large number of man-created failures and often by an even larger number of failures which never existed at all.

Now let us consider the use of missiles, i.e., the actual flight. Figure 1 shows a representative plot of the probability of survival of a missile, as a function of flight time. Zero on the time scale indicates the instant of signal to launch. Electrical and other stresses have usually been applied before launch is desired. As a consequence, some missiles do not launch, as you can see. The missile represented here is rocket-launched, as mentioned earlier.

Shortly after launch, part of the guidance system is made operative. It is then discovered that more missiles have succumbed, prior to this instant. A while later, the rest of the guidance system is cut in and this act reveals another set of failures. From here on the failures occur at a much lower rate. In fact they occur at a rate approximating the failure rate during tests on the ground. The failure rate during the first portion of flight, on the other hand, is orders of magnitude larger than the ultimate rate.

The solid curve of the figure is representative of actual data and forms the upper bound on survival probability. What is probably happening is indicated by the dotted curve. The failures cannot be recognized until the relevant circuit is put to work.

The figure illustrates that the big gain to be made in missile reliability comes from reducing the early, extremely high failure rate. The fact that the rate is so high indicates that the environment is much worse

at launch than on the test bench. This flight environment is probably the biggest unknown in any missile program.

In order to determine the flight environment and which components fail therein, several courses of action are open. Recovery of missiles seems the simplest, at first glance. We do now recover many missiles: these are the ones which did not fail, however. What we need is to recover those which do fail. Another approach is to use telemetering. However, with hundreds of components and thousands of connections to monitor, telemetering is really inadequate. In addition, the interaction between the telemetering and normal missile functions would have to be considered. Telemetering has been used in technical and tactical trials, of course, and found to be very useful as a diagnostic aid, even with the limited information sent back. A third possibility is to use a rocket sled. The sled would certainly produce the mechanical environment but has the defect

that flight pressure and temperature cannot be reproduced.

In summary, there exist schemes for reporting and acting upon chance component failures in conventional electronic equipment. Use of such schemes to improve bench-test reliability of guided missiles requires recognition that the failure data are diluted. Many failures are due not to the nature of the components involved but result from the use and abuse of the guided missile in which they are employed. The records are further diluted by reports of failures which never occurred at all.

Further, the real gain to be made in missile reliability requires investigation of early in-flight failures. We now know neither which components fail at this time, nor the magnitude of the stresses involved. In order to improve flight reliability we need information on both points. Such information cannot be gleaned from current failure-recording systems related to preflight operation of the missiles.

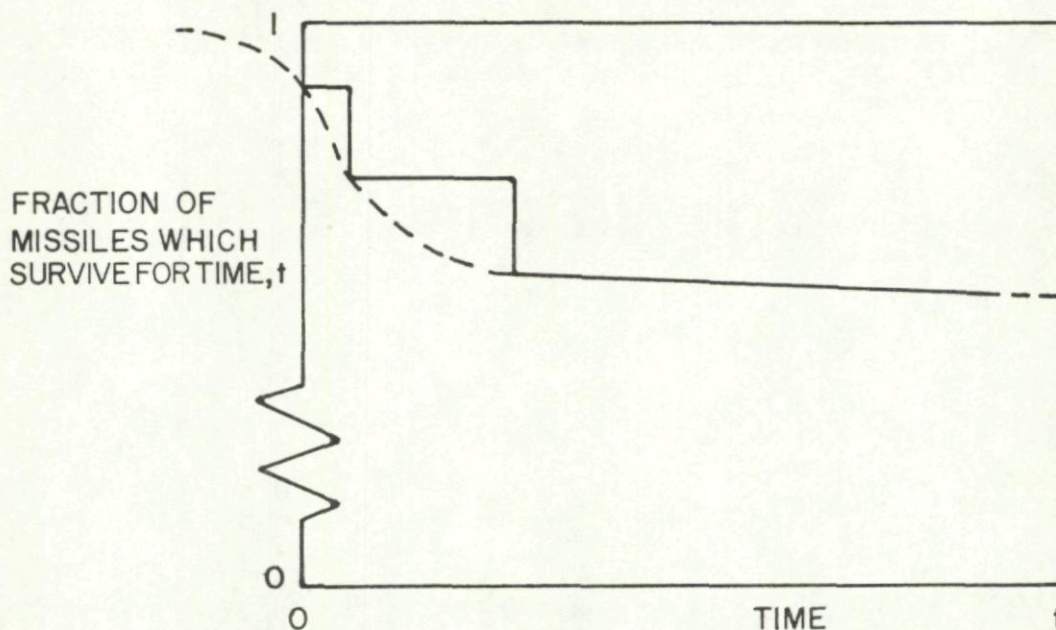


Fig. 1. Fraction of missiles which survive for time of flight, t , for a representative missile.

DISTRIBUTION

Copies of AGARD publications may be obtained in the various countries at the addresses given herewith.

BELGIUM	Centre National d'Etudes et de Recherches Aéronautiques 11, rue d'Egmont, Bruxelles, Belgique
CANADA	Director of Scientific Information Service Defence Research Board Department of National Defence "A" Building, Ottawa, Ontario, Canada
DENMARK	Military Research Board Defence Staff Kastellet, Copenhagen Ø, Denmark
FRANCE	ONERA (Direction) 25, avenue de la Division-Leclerc Châtillon-sous-Bagneux, Seine, France
GERMANY	Wissenschaftliche Gessellschaft fur Luftfahrt Zentralstelle der Luftfahrtdokumentation Munchen 64, Flughafen, Germany Attention: Dr. Ing. H. J. Rautenberg
GREECE	Greek Nat. Def. Gen. Staff B. MEO Athens, Greece
ICELAND	Iceland Delegation to NATO Palais de Chaillot Paris 16, France
ITALY	Centro Consultivo Studi e Ricerche Ministero Difesa Aeronautica Via Salaria 336, Roma, Italy
LUXEMBOURG	

NETHERLANDS

Netherlands Delegation to AGARD
Kanaalstraat 10
Delft, Holland

NORWAY

Chief Engineering Division
Royal Norwegian Air Force
Deputy Chief of Staff/Materiel
Myntgaten 2
Oslo, Norway
Attention: Lt. Col. S. Heglund

PORTUGAL

Subsecretariado da Estado da Aeronautica
Av. da Liberdade 252
Lisbon, Portugal
Attention: Lt. Col. Jose Pereira do Nascimento

TURKEY

M. M. Vekaleti
Erkaniharbiyei Umumiye Riyaseti
Ilmi Istisare Kurulu Mudurlugu
Ankara, Turkey
Attention: Colonel Fuat Ulug

UNITED KINGDOM

Ministry of Supply
T.I.L., Room 009A
First Avenue House
High Holborn
London W.C.1, England

UNITED STATES

National Advisory Committee for Aeronautics
1512 H Street, N. W.
Washington 25, D. C., U. S. A.



*Printed by Technical Editing and Reproduction Ltd
95 Great Portland St. London, W.1.*

139197